




# Bisociative Literature-Based Discovery: Lessons Learned and New Word Embedding Approach

Nada Lavrač<sup>1,2</sup> · Matej Martinc<sup>1,3</sup> · Senja Pollak<sup>1</sup> · Maruša Pompe Novak<sup>4</sup> · Bojan Cestnik<sup>1,5</sup> 

Received: 21 March 2020 / Accepted: 8 September 2020 / Published online: 6 October 2020  
© The Author(s) 2020

## Abstract

The field of bisociative literature-based discovery aims at mining scientific literature to reveal yet uncovered connections between different fields of specialization. This paper outlines several outlier-based literature mining approaches to bridging term detection and the lessons learned from selected biomedical literature-based discovery applications. The paper addresses also new prospects in bisociative literature-based discovery, proposing an advanced embeddings-based technology for cross-domain literature mining.

**Keywords** Literature-based discovery · Cross-domain bisociations · Computational creativity · Embeddings technology

---

✉ Bojan Cestnik  
bojan.cestnik@temida.si

Nada Lavrač  
nada.lavrac@ijs.si

Matej Martinc  
matej.martinc@ijs.si

Senja Pollak  
senja.pollak@ijs.si

Maruša Pompe Novak  
Marusa.Pompe.Novak@nib.si

<sup>1</sup> Jožef Stefan Institute, Jamova 39, 1000 Ljubljana, Slovenia

<sup>2</sup> University of Nova Gorica, Vipavska 13, 5000 Nova Gorica, Slovenia

<sup>3</sup> Jožef Stefan International Postgraduate School, Jamova 39, 1000 Ljubljana, Slovenia

<sup>4</sup> National Institute of Biology, Večna pot 111, 1000 Ljubljana, Slovenia

<sup>5</sup> Temida d.o.o., Dunajska cesta 51, 1000 Ljubljana, Slovenia

## Introduction

Growing amounts of available knowledge and data exceed human analytic capabilities. Therefore, new technologies that help analyzing and extracting useful information from large amounts of data need to be developed and used for analytic purposes. Understanding complex phenomena and solving difficult problems often require 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, and in particular association rule learning [2], scientific discovery requires creative thinking to connect seemingly unrelated information, for example, using metaphors or analogies between concepts from different domains. These kinds of context crossing associations, called bisociations [19], are often needed for innovative discoveries.

This paper addresses a computational creativity task of bisociative knowledge discovery from scientific literature that we name bisociative literature-based discovery. This task is at the intersection of two research areas: literature-based discovery [6] and bisociative knowledge discovery [3], which are briefly introduced below.

In literature-based discovery (LBD) [6]—and in particular in cross-domain literature mining that addresses knowledge discovery from two (or more) 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 and Smalheiser [37, 40] developed early LBD approaches to detecting interesting b-terms to uncover the possible cross-domain relations among previously unrelated concepts. Their approach, known as the ‘ABC model of knowledge discovery’, addresses the so-called closed discovery setting [43], 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 (b-terms) in  $B$  to validate the hypothesized connection between  $A$  and  $C$ .

Similarly, bisociative knowledge discovery [3] addresses a data mining task where two (or more) domains of interest are searched for bridging concepts (bridging terms or links). Using either the same representation of different domains or different representations of the same domain, bridging concepts can be detected either as nodes bridging different graphs, as subgraphs linking different graphs, as bridging links in terms of graph similarity, or as bridging terms appearing in separate document corpora, which is referred to as bridging term discovery in this paper.

Until recently, literature-based discovery and bisociative knowledge discovery approaches to cross-domain literature mining used conventional bag of words (BoW) vector representation of text, using term frequency inverse document frequency (TF-IDF) word weighting heuristics. Recent text-mining approaches started exploiting neural networks-based text representations, using text embedding methods that use large corpora of documents to extract numeric vector

representations for words, sentences, and/or documents. In this paper, we exploit the power of word embeddings [25, 27], which refer to vector representations of words, where each word is assigned a vector of several hundred dimensions in the transformed  $n$ -dimensional numeric vector space. Embedding approaches have started emerging also in the area of computational creativity [1, 10] and literature-based discovery [24].

The contributions of this paper are many-fold. The paper first reflects on the lessons learned from our past research in cross-domain literature mining,<sup>1</sup> focusing on outlier document detection as means for more effectively searching for novel bridging terms. Second, we propose an embedding-inspired conceptual framework for creative bisociative LBD, based on a novel concept of bridging by relational bisociation. Third, we propose a new bisociative LBD methodology, using word embeddings for relational bisociation discovery. Finally, we show-case the potential utility of this approach on a new biological research problem of finding connections between circadian rhythm and plant defense domains, where the results of this proof of concept evaluation indicate that the new methodology is very relevant for LBD research.

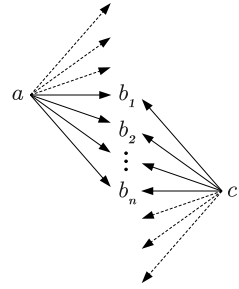
The paper is structured as follows. “[Background and related work](#)” presents the related work in literature-based discovery (LBD) and bisociative knowledge discovery, including the previously published relationship between the two [21, 31]. It presents also the related work in representation learning using the embedding technology. “[Past LBD results and lessons learned](#)” outlines selected approaches to cross-domain literature mining via outlier document detection and exploration [31, 36], together with the lessons learned from this research. “[Towards creative embeddings-based bisociative LBD](#)” proposes a novel creative discovery research direction based on the recent word embedding technology, with a proof of concept experiment in a biological domain, together with the lessons learned from this LBD application. Finally, “[Conclusions and further work](#)” concludes with a summary and plans for further research.

## Background and Related Work

This section presents the related work. “[Literature-based discovery](#)” introduces literature-based discovery (LBD), which is the main topic of this research. “[Bisociative knowledge discovery](#)” presents the area of computational creativity named bisociative knowledge discovery and the connection between bisociative knowledge discovery and LBD, as published in our past research [21, 31]. Finally, “[Embeddings](#)” briefly introduces embeddings, the contemporary representation learning technology resulting from recent research in neural networks, which is the enabler for the proposed embedding-based bisociative LBD methodology introduced in “[Towards creative embeddings-based bisociative LBD](#)”.

<sup>1</sup> These lessons have been published also in the ICCG-2020 paper by Lavrač et al. [22].

**Fig. 1** Closed discovery process defined by Weeber et al. [43]



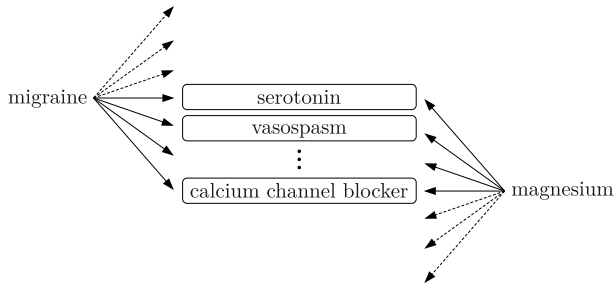
## Literature-Based Discovery

In literature-based discovery (LBD) [6]—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) that carry the potential of revealing the links connecting the separate domains.

Early work in LBD [37, 40] developed approaches to assist the user in literature-based discovery by detecting interesting cross-domain terms with a goal to uncover the possible relations between previously unrelated concepts. The ARROWSMITH online system, developed by Smalheiser and Swanson [37], 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.<sup>2</sup> Their approach, known as the ‘ABC model of knowledge discovery’, addresses several settings, including the closed discovery setting [43], 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 concept (term)  $b$  in  $B$  to support the validation of the hypothesized connection between  $A$  and  $C$ . The closed discovery setting, which is the most frequently addressed LBD setting, is illustrated in Fig. 1.

Swanson’s seminal work has shown that databases such as PubMed can serve as a rich source of yet hidden relations between usually unrelated topics, potentially leading to novel insights and discoveries. By studying two separate literatures, i.e., the literature on migraine headache and the articles on magnesium, Swanson [39] discovered ‘Eleven neglected connections’, all of them supportive for the hypothesis that magnesium deficiency might cause migraine headache. Figure 2 illustrates the closed discovery setting on the Swanson’s task of finding the terms linking the ‘migraine’ and ‘magnesium’ domains. Swanson’s literature mining results have been later confirmed by laboratory and clinical investigations. This well-known example

<sup>2</sup> 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.



**Fig. 2** Closed discovery when exploring migraine and magnesium documents, with b-terms identified by Swanson et al. [41]

has become the gold standard in the literature mining field and has been used as a benchmark in several studies [17, 23, 38, 43].

Inspired by this early work, literature mining approaches were further developed and successfully applied to different problems, such as finding associations between genes and diseases [16], diseases and chemicals [44], and others. Supporting the user in effectively searching for bridging terms (b-terms) provided a motivation for developing the CrossBee approach to bridging term detection applicable in the closed discovery setting [17], implemented through ensemble-based term ranking, where an ensemble heuristic composed of six elementary heuristics was constructed for term evaluation.

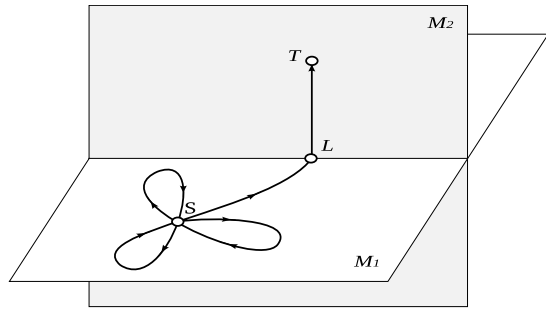
The work of Kastrin et al. [18] is complementary to other LBD approaches, as it uses different similarity measures (such as common neighbors, Jaccard index, and preferential attachment) for link prediction of implicit relationships in the Semantic MEDLINE network. Holzinger et al. [15] describe several 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. A comprehensive survey of modern literature-based discovery approaches in biomedical domain can be found in [13, 33].

Our past research [31, 36] suggests that bridging terms are more frequent in documents that are in some sense different from the majority of documents in a given domain. For example, Sluban et al. [36] have shown that such documents, considered being outlier documents of their own domain, contain a substantially larger amount of bridging/linking terms than the regular non-outlier documents. This approach, using the OntoGen tool [12], is described in some detail in “[Past LBD results and lessons learned](#)”.

## Bisociative Knowledge Discovery

Bisociative knowledge discovery is a challenging task motivated by a trend of over-specialization in the research and development, which usually results in deep and relatively isolated silos of knowledge. Scientific literature too often remains closed and cited only in professional subcommunities. The information that is related across different contexts is difficult to identify using associative approaches, like

**Fig. 3** Koestler's schema of bisociative discovery in science [19, p. 107], illustrating the creative act of finding links (from  $S$  to target  $T$ ) that lead 'out-of-the-plane' via intermediate, bridging concepts ( $L$ )



the standard association rule learning [2] known from the data mining and machine learning literature. Therefore, the ability of literature mining methods and software tools to support the experts in their knowledge discovery processes—especially in searching for yet unexplored connections between different domains—is becoming increasingly important.

Arthur Koestler [19] argued that the essence of creativity lies in “perceiving of a situation or idea . . . in two self-consistent but habitually incompatible frames of reference”, and introduced the expression bisociation to characterize this creative act. More specifically, Koestler's notion of bisociation was originally defined as follows.

“The pattern ... is the perceiving of a situation or idea,  $L$ , in two self-consistent but habitually incompatible frames of reference,  $M_1$  and  $M_2$ . The event  $L$ , in which the two intersect, is made to vibrate simultaneously on two different wavelengths, as it were. While this unusual situation lasts,  $L$  is not merely linked to one associative context but bisociated with two.”

Koestler found bisociation to be the basis for human creativity in seemingly diverse human endeavors, such as humor, science, and arts. The concept of bisociation is illustrated in Fig. 3. It should be noted that context crossing is subjective, since the user has to move from his ‘normal’ context (frame of reference) to an *habitually incompatible context* to find the bisociative link. In Koestler's terms (Fig. 3), a habitual frame of reference (plane  $M_1$ ) corresponds to the domain defined by the user. Other domains represents different, habitually incompatible contexts (in general, there may be several planes  $M_2$ ), where the creative act is to find links that lead ‘out-of-the-plane’ via intermediate, bridging concepts. Thus, contextualization and link discovery are two of the fundamental mechanisms in bisociative reasoning.

In summary, according to Koestler [19], bisociative thinking occurs when a problem, idea, event, or situation is perceived simultaneously in two or more ‘matrices of thought’ or domains. When two matrices of thought interact with each other, the result is either their fusion in a novel intellectual synthesis or their confrontation in a new aesthetic experience. Koestler regarded many different mental phenomena that are based on comparison (such as analogies, metaphors, jokes, identification, and anthropomorphism) as special cases of bisociation.

More recently, this work was followed by the researchers interested in the so-called bisociative knowledge discovery, where—according to [3]—two concepts are

**Table 1** Unifying Koestler’s and Swanson’s models of creative knowledge discovery [21, 31]

Koestler’s model	Swanson’s model
Bisociative link discovery process	Closed discovery process
Frames of reference (contexts) $M_1$ and $M_2$	Domains of interest $A$ and $C$
Bisociative cross-context link $L \in M_1 \cap M_2$	Bridging term $b \in \text{terms}(A) \cap \text{terms}(C)$

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. Bisociative knowledge discovery has become a topic of extensive research, addressing the discovery of bridging links or bridging concepts crossing between different domains and representations.

In conclusion, let us summarize the previously published [21, 31] relationship between bisociative knowledge discovery and Swanson’s ABC model for literature-based discovery, where the particular focus of interest is the relationship between Koestler’s bisociative link discovery framework and Weeber’s closed discovery framework, as summarized in Table 1. Similar to a bisociation, which is according to Koestler a result of processes of mind when making new associations between concepts  $S$  and  $T$  from usually separated contexts (illustrated in Fig. 3), literature-based discoveries in Swanson’s ABC model are a result of uncovering links between concepts  $a$  and  $c$  from disjoint literatures  $A$  and  $C$  (illustrated in Fig. 1). In terms of Koestler’s model, the two domains  $A$  and  $C$ , investigated in the closed literature-based discovery framework, correspond to the two habitually incompatible frames of reference,  $M_1$  and  $M_2$ . Moreover, the bridging terms  $b_1, b_2, \dots, b_n$  that are common to literature  $A$  and  $C$  clearly correspond to Koestler’s notion of a situation or idea,  $L$ , which is not merely linked to one associative context, but bisociated with two contexts  $M_1$  and  $M_2$ .

## Embeddings

In terms of representation learning, our past LBD research that led to the lessons learned described in “[Past LBD results and lessons learned](#)” was based on using the standard TF-IDF weighted BoW vector representations of text documents [7, 17, 31, 36]. On the other hand, the novel LBD methodology proposed in this paper in “[Towards creative embeddings-based bisociative LBD](#)” exploits contemporary representations of text documents using embeddings, given that current research in natural language processing demonstrates that representation learning using embeddings is much more effective than using the standard TF-IDF BoW vector representation. The embedding approach to representation learning can be defined as follows.

**Embeddings** Given input data of a given data type and format, find a tabular representation of the data, where each row represents a single data

instance, and each column represents one of the dimensions in the  $d$ -dimensional numeric vector space  $\mathbb{R}^d$ .

The embedding technology is a prominent side effect of the recent revival of neural networks (NN), in which the information is represented by activation patterns in interconnected networks of primitive units (neurons). This enables concepts to be gradually learned by an NN from the observed data by modifying the connection weights between the hierarchically organized units. These weights that can be extracted from neural networks can be used as a spatial representation that transforms relations between observed entities (data instances) into distances.

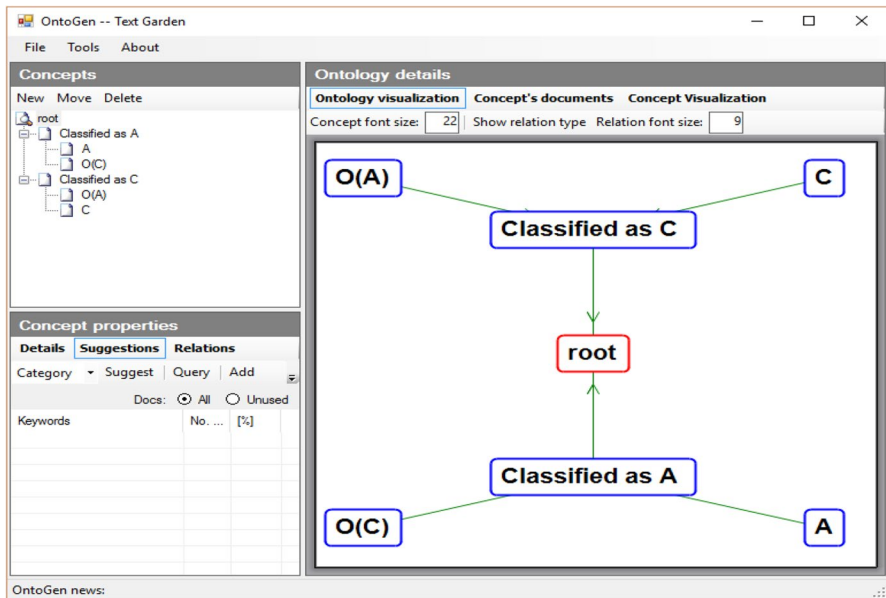
Recently, the embedding approach became a prevalent way to build representations for many different types of entities, e.g., graphs, electronic health records, images, relations, recommendations, as well as texts (documents, sentences and/or words). Word embeddings [25, 27], which are in the focus of our research described in “Towards creative embeddings-based bisociative LBD”, use large corpora of documents to extract vector representations of words, assigning each word a vector of several hundred dimensions. The first neural word embeddings like word2vec [25] produced one vector for each word, irrespective of its polysemy (e.g., for a polysemous word like bank, word2vec produces a single representation vector, and ignores the fact that bank can present both a financial institution and a land sloping down to a water mass). Recent developments like ELMo [30] and BERT [9] take a context of a sentence into account and produce different word vectors for different contexts of each word. A further improvement of neural word embeddings for texts uses multi-task prediction (inclusion of several related textual prediction tasks).

## Past LBD 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 by Petrič et al. [31] and Sluban et al. [36] 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. [31] uses document clustering to find outlier documents. The approach consists of two steps. In the first step, the OntoGen clustering algorithm by Fortuna et al. [12] is applied to cluster the merged document set  $A \cup C$ , consisting of documents from two domains  $A$  and  $C$ . The result of unsupervised clustering is two document clusters:  $A' = \text{Classified as } A$  (i.e., documents from  $A \cup C$  classified as  $A$ ), and  $C' = \text{Classified as } C$  (i.e., documents from  $A \cup C$  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, as illustrated in Fig. 4.





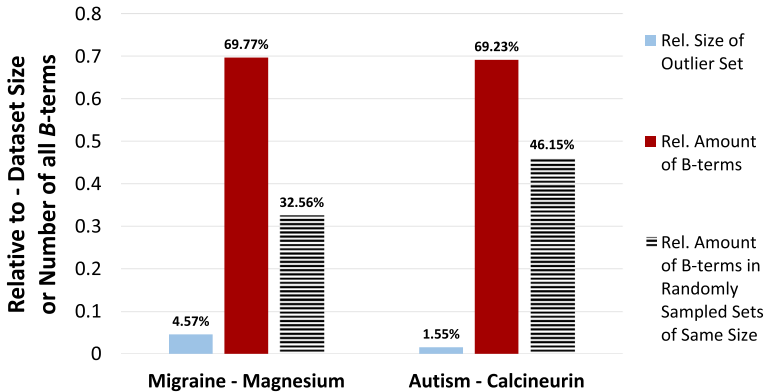
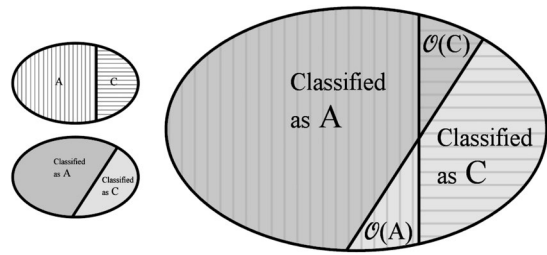
**Fig. 4** Target domain documents from literatures *A* and *C*, clustered according to the OntoGen's two-step approach, first using unsupervised and then supervised clustering to obtain outlier documents  $O(A)$  and  $O(C)$  of literatures *A* and *C*, respectively. The figure illustrates the outlier detection approach implemented using OntoGen, addressing the outlier detection framework that is conceptually explained in Fig. 5

### 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 [41] and autism–calcineurin [32] domain pair datasets, which have lists of concept bridging terms (b-terms) confirmed by the medical experts. The experimental results obtained 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. [36], exploring a classification filtering approach to outlier detection, which was tested on the same domain pair data sets, migraine–magnesium [41] and autism–calcineurin [32] domain, which have lists of bridging terms (b-terms) confirmed by the medical experts. Sluban et al. [36] proposed to detect outlier documents using

**Fig. 5** Detecting outliers of a domain pair dataset  $A \cup C$ , using a document classification approach by Sluban et al. [36]

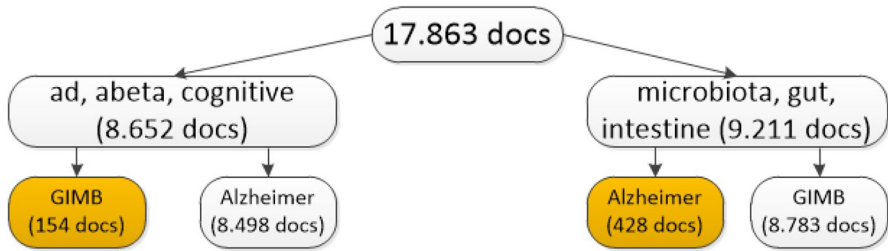


**Fig. 6** Presence of b-terms in the detected outlier sets of two domain pair datasets

classification algorithms for classification noise filtering, first suggested by Brodley and Friedl [5]. Having documents from two domains of interest  $A$  and  $C$ , Sluban et al. [36] 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 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  $O(C)$ , as illustrated in Fig. 5.

The experimental results obtained by Sluban et al. [36] 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 Fig. 6.

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



**Fig. 7** Two-level cluster hierarchy constructed with ontoGen from the dataset of 17,863 papers in the Alzheimer’s disease–gut microbiome domain pair

much smaller subset of documents, where a great majority of b-terms are present and more frequent.

When applying OntoGen on the documents of the new application domain using the Alzheimer’s disease–gut microbiome domain pair [7], 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, and intestine*. Four second-level subclusters separate documents according to their original search keywords for Alzheimer’s disease and gut microbiome, as illustrated in Fig. 7.

### Lesson Learned 2: Excluding intersecting documents

In Alzheimer’s 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,<sup>3</sup>

<sup>3</sup> 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 was violated, the methodology would fail due to biased heuristics calculations.

### Lesson Learned 3: Selecting only outlier documents

resulting in 17,863 documents kept in the  $A \cup C$  document set used for further exploration.

The hypothesis that the search for bridging terms can be reduced to manageable subsets of documents was confirmed in our experiments. In the Alzheimer's 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 [17] to extract 4723 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. 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<sup>4</sup> 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 4723 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

---

<sup>4</sup> Elsa Fabretti.

from relevant b-term discoveries confirmed by the domain expert [7].

Compared to outlier document detection using OntoGen, an upgraded methodology proposed by Cestnik et al. [7] was implemented in a reusable outlier-based LBD methodology in a web-based text-mining platform TextFlows<sup>5</sup> [29] that allowed us to construct and execute advanced text-mining workflows. The workflow shown in Fig. 8 consists of seven steps implemented as subprocesses. The connections between subprocesses represent the flow of documents from one subprocess to another. In overview, steps 1–3 represent the outlier detection part, and steps 4–7 represent cross-domain exploration for b-term detection.

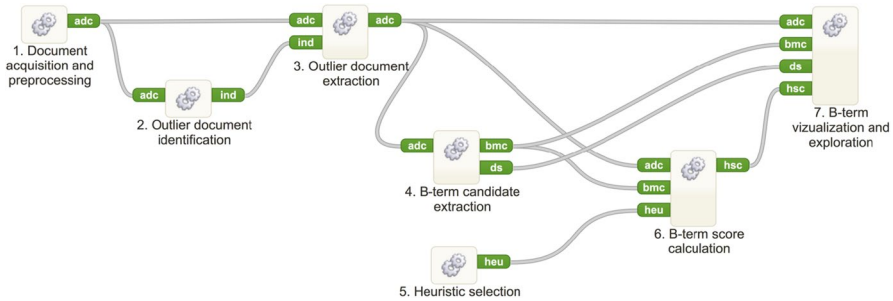
#### Lesson Learned 5: TextFlows workflow helping experts

In the experiments using the TextFlows workflow, the NoiseRank ensemble-based outlier detection approach [35] 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 Fig. 8, components that constitute the CrossBee HCI interface [17] 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.

## Towards Creative Embeddings-Based Bisociative LBD

In this section, we first formally define bisociation and the specific bisociative patterns that are searched for bisociative knowledge discovery (i.e., bridging concepts, bridging graphs, and bridging by structural similarity), including the novel concept of bridging by relational bisociation in “[Formal framework for creative bisociative LBD](#)”. The potential of the embeddings technology for creative knowledge discovery is explained in “[Word embeddings potential for creative knowledge discovery](#)”. “[Novel embeddings-based bisociative LBD methodology](#)” presents the proposed word embeddings-based bisociative LBD methodology, and explores the creativity

<sup>5</sup> <http://textflows.org>.



**Fig. 8** A top-level workflow of the LBD methodology in TextFlows [29]

potential of word embeddings in an LBD closed discovery setting, assuming an expert-defined relationship of interest between two terms  $a_1$  and  $a_2$  in domain  $A$  and an unknown relationship to be discovered for a given seed concept  $c$  and an unknown/yet to be discovered term  $x$  in domain  $C$ . “[Embeddings-based relational LBD experiment conducted on the circadian rhythm and plant defense domains](#)” briefly outlines the experimental setting of the experiments conducted on the circadian rhythm–plant defense domain pair, where a proof-of-concept result evaluation is given in “[Results](#)”. This section concludes by a summary and lessons learned from these experiments in “[Summary and lesson learned from these experiments](#)”.

## Formal Framework for Creative Bisociative LBD

Bisociation is essentially a creative endeavor. To connect pieces of information from previously unrelated domains, a person must activate some form of creative mechanism. This creative aspect is what allows one to go beyond one-dimensional associations. This has been recognized in several psycho-cognitive theories related to creativity, which share the principle that a strong connection exists between creative activity and the ability to establish relations between seemingly unrelated domains.

Divergent reasoning can be achieved—to a certain degree—by means of cross-domain exploration in multi-domain databases. Such a model must provide mechanisms for mapping concepts and transferring meanings. According to Koestler [19], in addition to metaphor, the well-known examples of mechanisms that can be used in cross-domain knowledge transfer are analogy and bisociation. Before addressing bisociative computational creativity, we continue with the presentation of a formal definition of bisociation, as formulated by Dubitzky [11].

**Definition 1** (Domain theory) A domain theory  $D_i$  defines a set of concepts (knowledge units) that are associated with a particular domain  $i$ .

**Definition 2** (Knowledge base) A knowledge base  $K_i$  is defined as a subset of a domain theory  $D_i$ ; that is,  $K_i \subseteq D_i$ .

$D_i$  denotes a domain theory which represents the total knowledge within a domain. The union of all domains then represents the universe of discourse:  $\cup_i D_i = U$ . Many domain theories overlap:  $\exists i, j : D_i \cap D_j \neq \emptyset$ . Let  $U$  denote the universe of discourse, which consists of all concepts. Let  $c \in U$  denote a concept in  $U$ . Within  $U$ , a problem, idea, situation, or event  $\pi$  is associated with concepts  $X \subset U$ . Typically, a subset  $P \subset X$  is used to reason about  $\pi$ .

Let  $R$  denote a reference system or intelligent agent which possesses exactly one knowledge base (empty or non-empty) per domain theory  $D_i$ .  $K_i^R \in D_i$  denotes the knowledge base with respect to  $R$  and  $D_i$ .

$K^R = \cup_i K_i^R$  denotes the entire set of  $K$  incorporated in the reference system of  $R$ .  $K^R$  represents the total knowledge that  $R$  has in all the domains. For example,  $R$  may have non-empty knowledge bases for chess, but an empty one for geometry.

**Definition 3** (Association) Let  $\pi$  denote a concrete problem, situation of event and let  $X \subset U$  denote the concepts associated with  $\pi$ . Furthermore, let  $K_i^R$  denote an agent-specific knowledge base. Association occurs when elements of  $X$  are active or perceived in  $K_i^R$  at time  $t$  only.

For example, at time  $t$ , the concepts  $A = \{c_1, c_2, c_3\}$  may be active in  $K_i^R$  only. In this case, we say that the concepts in  $A$  are associated.

**Definition 4** (Habitually incompatible knowledge bases) Two agent-specific knowledge bases  $K_i^R$  and  $K_j^R$  ( $i \neq j$ ) are habitually incompatible if, at a given point in time  $t$ , there is no concept  $c : c \in K_i^R \wedge c \in K_j^R$  that is active or perceived simultaneously in  $K_i^R$  and  $K_j^R$ .

**Definition 5** (Bisociation) Let  $\pi$  denote a concrete problem, situation or event, and let  $X \subset U$  denote the concepts associated with  $\pi$ . Furthermore, let  $K_i^R$  and  $K_j^R$  be such that  $i \neq j$ . Bisociation occurs when elements of  $X$  are active or perceived simultaneously in both  $K_i^R$  and  $K_j^R$  at a given point in time  $i$ .

For example, at time  $t$ , the concepts  $B = \{c_1, c_2, c_3\}$  may be active or perceived simultaneously in  $K_i^R$  and  $K_j^R$ . In this case, the concepts in  $B$  are bisociated.

Bisociation cannot be equated with creativity in general. It is instead a special case of combinatorial creativity, which refers to novel combinations of familiar ideas: the creative aspect here is in the discovery of previously non-existing connections between domains, especially if each of the domains, or the elements repurposed from each, are very familiar. As put by Koestler [19], “the more familiar the parts, the more striking the new whole”. This is so because creation is never really a de novo nor random activity; it requires meaningful combination of elements.

Starting from Kostler’s [19] concept of bisociation, concrete bisociative patterns that are searched for in bisociative knowledge discovery include: bridging concepts, bridging graphs, and bridging by structural similarity [20]:

Bridging concepts	This is the most natural type of bisociation: a concept connecting two domains. In practice, different literatures from different domains are explored, and some terms connecting the two are found. This is the kind of pattern originally explored by Swanson. These connecting terms allow us to corroborate hypotheses linking the two domains. Bridging concept in the intersection of two domains <i>A</i> and <i>C</i> is illustrated in Fig. 9.
Bridging graphs	More complex bisociations are modeled by bridging graphs, in a network representation. This is similar to bridging concepts, but in this case, what connects two different domains is a subset of related concepts.
Bridging by structural similarity	This is the most complex kind of bisociation, whereby, again in a network representation, subsets of concepts in each domain share structural similarities, illustrated in Fig. 10.

Bisociations based on structural similarity are represented by relations and/or subgraphs of two different, structurally similar domains [20], as illustrated in Fig. 10. This type of bisociation is according to [20] the most abstract pattern with the potential for new cross-domain discoveries, which, e.g., vertex similarity methods can identify.

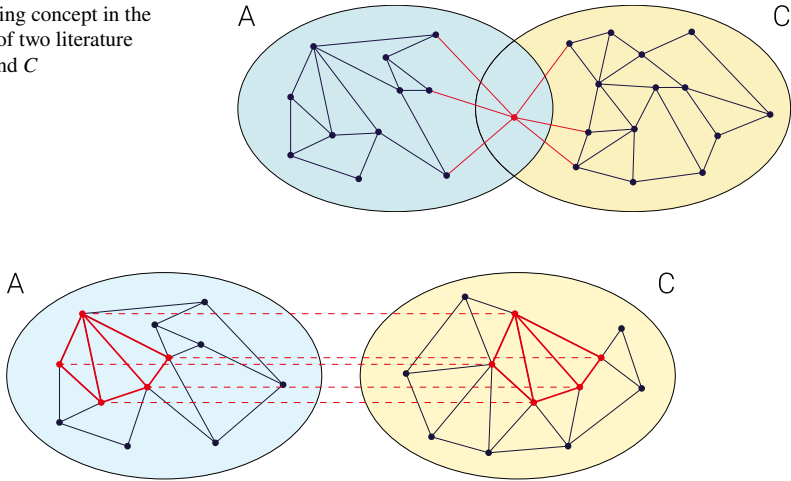
A special case of bridging by structural similarity is the concept of bridging by relational bisociation, as illustrated in Fig. 11, which will be explained and used in the novel methodology proposed in “[Novel embeddings-based bisociative LBD methodology](#)”.

## Word Embedding Potential for Creative Knowledge Discovery

Note that in this research, we neither use the TF-IDF representation of documents nor do we use document embeddings; instead, we focus on 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 [25], GloVe [28], or FastText [4], 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 [27]. A well-known example, illustrating this notion, is that word embeddings may explicitly find relations between words, as well as discover analogies between word pairs, such as that, e.g., the relation between Madrid and Spain is very similar to that between Paris and France in the embedded vector space (see Fig. 12).



**Fig. 9** Bridging concept in the intersection of two literature domains A and C



**Fig. 10** Bridging by structural similarity of graphs [20]

Note that the analogies can be discovered within a single domain, as illustrated in Fig. 12. On the other hand, research in cross-lingual embeddings [8] has demonstrated the ability of aligning embeddings spaces across languages, which can be used as a basis for finding analogies across corpora in different languages [42], as investigated in the current EMBEDDIA EU project.<sup>6</sup>

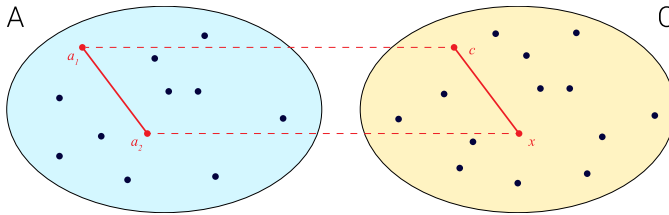
In this paper, we propose a novel methodology, based on the idea of translating the cross-lingual setting to a cross-domain setting: instead of considering two different languages, we consider two separated domains A and C, use contemporary alignment methods [8] to align related concepts in the two domains, and finally perform analogy detection across the two domains [42]. In this way, we find bisociations by implementing the idea of bridging by relational bisociation.

### Novel Embedding-Based Bisociative LBD Methodology

Most important for this paper is the property of word embeddings that they can capture regularities beyond simple relatedness, such as analogies [27], illustrated in Fig. 12. In the particular closed literature-based discovery setting of interest to this research, we implement the concept of bridging by relational bisociation.

**Bridging by relational bisociation** We propose a particular setting of bridging by relational bisociation, illustrated in Fig. 11, where we are interested whether given a specific relation between two concepts  $a_1$  and  $a_2$

<sup>6</sup> [www.embeddia.eu](http://www.embeddia.eu), see details in Acknowledgements.



**Fig. 11** Bridging by relational bisociation, the concept newly introduced in this paper

in first domain  $A$ , one can bisociatively discover an analogous relation between concepts  $x$  and  $c$  in second domain  $C$ , where  $c$  is a given concept and  $x$  is a new concept that we are trying to find. More formally, this can be written in the form of an analogy (i.e., bisociation) between two separate domains  $A$  and  $C$  as follows:

$$a_1 \text{ rel } a_2 == x \text{ rel } c.$$

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

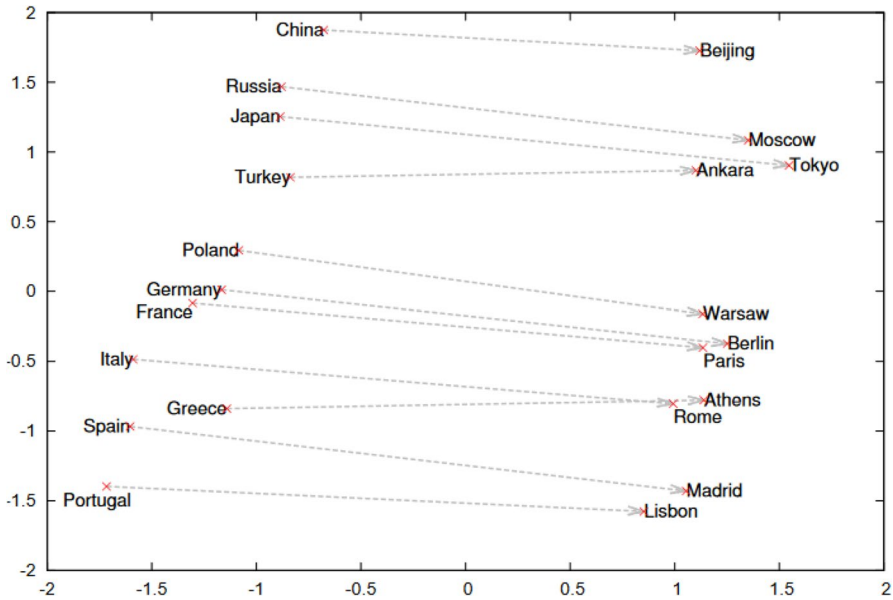
$$x = \text{emb}(a_1) + \text{emb}(a_2) - \text{emb}(c).$$

Finally, once  $x$  is calculated, we need to find a set of concepts from the second domain  $C$  that have an embeddings representation most similar to  $x$  according to some predefined distance measure (e.g., the cosine similarity).

#### Methodology of bridging by relational bisociation

Proposed embedding-based bisociative LBD methodology for creative discovery of bisociated relationships between two domains  $A$  and  $C$  consists of the following steps:

1. Select two domains  $A$  and  $C$ , i.e., two document corpora such as circadian rhythm and plant defense, respectively.
2. Train separate word embeddings models for  $A$  and  $C$  to get  $\text{emb}(A)$  and  $\text{emb}(C)$ .
3. Perform alignment of  $\text{emb}(A)$  and  $\text{emb}(C)$  embeddings vector spaces.
4. Determine the relationships of interest in a given domain  $A$  between concepts  $a_1$  and  $a_2$  defined by the biology expert.
5. Perform the embeddings-based relational LBD with a known seed concept  $c$  in  $C$  by leveraging the ability of the embeddings representations to model analogy relations.
6. Evaluate a list of best-ranked relational bisociations.



**Fig. 12** Two-dimensional projection of embeddings illustrating capital–country relations. Picture taken from Mikolov [26]

### Embeddings-Based Relational LBD Experiment Conducted on the Circadian Rhythm and Plant Defense Domains

In this section, we report in detail on the experiments conducted on the circadian rhythm and plant defense domains. Our main goal was to identify potentially interesting new daily regulated mechanisms that are responsible for plant defence. Circadian rhythm in plants causes that some of their genes are expressed differently during the course of the day. Consequently, plants respond differently to disease-causing infection if they are infected at different times of the day (e.g., morning, noon, and evening). Therefore, one of the goals of our study was to identify new gene sets that are differently expressed in different parts of the day and are important for the defense of plants against the pathogen.

After obtaining 10,494 documents from PubMed containing article titles and abstracts (4346 from plant defence and 6148 from circadian rhythm), we replaced gene names with synonyms gathered in previous research projects (22,265 gene names mapped into 7863 synonyms). In addition, we pre-processed the documents to keep only gene-related terms (included in synonym list and from the gene dictionary containing additional 6083 gene names), which resulted in a substantial reduction of the input document corpus, which we called the genesOnly dataset. The experiments that were conducted following the methodology proposed in “[Novel embeddings-based bisociative LBD methodology](#)” served as a proof of concept to show that the new proposed embeddings-based methodology can be used for LBD.

On each of the two selected domains, circadian rhythm and plant defense, we trained a separate FastText embedding model [4]. FastText embeddings were chosen due to their ability to leverage both semantic and morphological information by representing each word as an average of its character  $n$  grams. This is useful in a setting with a relatively small domain corpora containing less semantic information, since morphological similarity in many cases translates to semantic relatedness. We used a skip-gram model with an embedding dimension of 100.

The resulting embedding models trained for each domain were in the next step aligned into a common vector space. We opted for a supervised alignment approach, which relies on a training dictionary of identical words from both domains that are used as anchor points to learn a mapping from the source to the target space with a Procrustes alignment [8]. Train and test dictionaries were constructed by taking 5000 most frequent words from both domains (i.e., words that appear in both domain and have the largest sum of frequencies) and then split randomly into a train dictionary containing two-thirds of the words (3333) and a test dictionary containing one-third of the words (1667).

The success of the alignment was measured on the test dictionary in terms of  $\text{precision@k}$ , where  $\text{precision@1}$  represents a share of model's correct alignments (exact matches) in a set of all alignments, and  $\text{precision@5}$  represents a share of model's alignments in a set of all alignments, where the correct match for the word is found in the set of 5 most probable alignments predicted by the model. In the conducted experiment, we report the  $\text{precision@1}$  of 0.4 and  $\text{precision@5}$  of 0.55.

Next, we asked a biology expert to identify a list of genes related to the circadian rhythm domain. The following list was produced:

1. CCA1 = CIRCADIAN CLOCK ASSOCIATED1
2. LHY = LATE ELONGATED HYPOCOTYL
3. TOC1 = TIMING OF CAB EXPRESSION 1
4. PRR1 = PSEUDO-RESPONSE REGULATOR 1
5. GI = GIGANTEA
6. LNK1 = NIGHT LIGHT-INDUCIBLE AND CLOCK-REGULATED 1
7. PRR5 = PSEUDO-RESPONSE REGULATOR 5
8. ELF4 = EARLY FLOWERING 4
9. PRR9 = PSEUDO-RESPONSE REGULATOR 9
10. PRR7 = PSEUDO-RESPONSE REGULATOR 7
11. PCL1 = PHYTOCLOCK 1
12. ELF3 = EARLY FLOWERING 3

In addition, we also took more general key concepts from the circadian rhythm domain:

13. NEGATIVE FEEDBACK LOOP
14. OSCILLATOR
15. CLOCK.

According to the methodology explained in “[Novel embeddings-based bisociative LBD methodology](#)”, we tried to identify a list of genes related to the concept of plant defense in the similar way the genes from the above list are related to the concept of circadian rhythm. First, we calculated embedding  $x$  according to the following equation:

$$x = \text{emb}(a1) + \text{emb}(a2) - \text{emb}(c),$$

where  $a1$  is a concept circadian rhythm,  $a2$  is a gene from the above list, and  $c$  is a concept plant defense.

Finally, once  $x$  was calculated for each of the genes from the above list, we searched for a set of concepts from the plant defense domain that have an embeddings representation most similar to  $x$  according to the cosine similarity. To limit the results only to genes or gene-related concepts, the concepts from the second domain were considered only if they appeared in the reduced genesOnly dataset. Ten genes or gene-related concepts with the representation most similar to each of the calculated  $x$ s were identified and given to the biology expert for the evaluation.

## Results

The biology domain expert evaluated the selected set of output terms for all given analogy inputs. More specifically, for the analogies— $a2$  is as important to  $a1$  (circadian rhythm) as  $x$  is to  $c$  (plant defense)—for each input relation, the resulting list of 10 candidates most similar to  $x$  (according to the cosine similarity between the candidate’s embedding and  $x$ ) were evaluated by the expert, who was given instructions to manually classify the relatedness between a candidate and the plant defense domain into the following four categories: NO, NOT AT ALL; NOT REALLY; MAYBE; YES. While YES is the category serving as a proof that the methodology works, MAYBE is the category containing very interesting terms from the knowledge discovery point of view, as, here, the experts might potentially search for novel knowledge.

First, we calculated the average precision at 10 ( $p@10$ ) for each output list of 10 candidates, as well as a microaveraged precision for the entire dataset (see Table 2). We can observe that the method performed very well. In 40% of the cases, the expert found in the scientific literature that the discovered relation between the plant defense concept and the proposed term  $x$  is meaningful. We can see that precision varies for different input relations, but the method was able to find at least one correct relation in the plant defense domain for each circadian rhythm input relation. For input relations between the concept circadian rhythm and genes ELF4, PRR9 and PRR7, six out of ten term candidates in the resulting candidate lists are related to the plant defense domain. On the other hand, the lowest results are for the input concept negative feedback loop, where only for one out of ten output terms, the expert found that the output term was relevant for the domain. A reason for this could be that the input term is one of the few terms, which is not a gene but rather a gene-related concept (text), and that it is a multi-word expression, for which the average embedding was first calculated (by averaging embeddings for each word in

the term) to obtain the term embedding, and, therefore, the results might be less precise.<sup>7</sup>

For the category MAYBE, which is the most interesting category for the new knowledge discovery and for which the outputs might possibly be investigated in detail in the future research by the domain experts, we can note that for all input relations but one, at least one out of 10 outputs was considered potentially interesting. In a knowledge discovery setting, where each discovery if resulting in new domain knowledge would have big impact, this was considered as a promising result.

As explained above, the biology expert evaluated 10 candidates for each input relation. These relations were ranked according to the cosine similarity between  $x$  and the candidate, with rank 1 representing the candidate closest to  $x$ , i.e., with the largest cosine similarity to  $x$ . Table 3 presents results for candidates with different ranks. Note that here we measured precision at 15, i.e., how many out of 15 predicted terms with a specific rank had been evaluated as related to the plant defense domain (P@15 yes) or as maybe being related to the plant defense domain (P@15 maybe). Interestingly, the correlation between precision at 15 and rank was not strong and better ranked candidates were not necessarily more correlated to the plant defense domain according to the evaluation. For example, the best evaluated candidates had rank 5, where P@15 yes was 0.6 and P@15 maybe was 0.133.

Next, we removed duplicate outputs, merged all the output terms from all the inputs, and calculated the class distribution for this list (see Table 4). The rationale for this procedure is that since, in our case, the  $a1$  and  $c$  were always the same (equivalent to domain names circadian rhythm and plant defence) and as the different  $a2$  all modeled the same relation— $a2$  is as important to  $a1$  (circadian rhythm) as  $x$  is important to  $c$  (plant defense)—we could treat also all results as a common list of relevant terms (genes). As the results indicate, about 37% of output terms were evaluated as relevant to the plant defense domain. Also, together with the category MAYBE, which indicates that the output is potentially relevant (but requires further research), this percentage of relevant terms increased to nearly 55%.

From 40 examples in the categories YES and MAYBE, 32 were gene names and 3 were proteins, while the rest referred to a disease or partial names of proteins, genes, etc. Below, we list ten terms and their full names that were classified in category YES and appeared in results of at least 3 input terms:

1. DMR1 = DOWNY MILDEW RESISTANT 1
2. CPR30 = CONSTITUTIVE EXPRESSER OF PR GENES 1
3. EIF4G = EUKARYOTIC TRANSLATION INITIATION FACTOR 4 G
4. SLAC1 = SLOW ANION CHANNEL-ASSOCIATED 1

<sup>7</sup> There are several possible multi-word expression aggregation approaches, such as summation of component word vectors, averaging of component word vectors, creating multi-word term vectors, etc. As comparing different techniques is beyond the scope of this study, we decided for the simple averaging technique, as the previous research on this topic conducted on the medical domain [14] found no statistically significant difference between any multi-word expression aggregation method.

**Table 2** Evaluation for 15 input relations

Source gene/term	No, not at all	Not really	Maybe	Yes	P@10 Maybe	P@10 Yes
CCA1	1	3	1	5	0.1	0.5
LHY	0	5	1	4	0.1	0.4
TOC1	0	1	3	6	0.3	0.6
PRR1	0	5	2	3	0.2	0.3
GI	1	5	2	2	0.2	0.2
LNK1	1	4	1	4	0.1	0.4
PRR5	0	3	3	4	0.3	0.4
ELF4	0	2	2	6	0.2	0.6
PRR9	0	4	0	6	0.0	0.6
PRR7	0	3	1	6	0.1	0.6
PCL1	2	6	0	2	0.0	0.2
ELF3	1	4	2	3	0.2	0.3
NEGATIVE FEED- BACK LOOP	7	0	2	1	0.2	0.1
OSCILLATOR	4	2	1	3	0.1	0.3
CLOCK	1	1	3	5	0.3	0.5
All	18	48	24	60	0.16	0.4

5. RFC3 = REPLICATION FACTOR C SUBUNIT 3
6. RTM1 = RESTRICTED TEV MOVEMENT 1
7. SNI1 = SUPPRESSOR OF NPR1-1
8. GRF6 = GROWTH-REGULATING FACTOR 6
9. NAC083 = NAC DOMAIN CONTAINING PROTEIN 83
10. XAP5 = XAP5 CIRCADIAN TIMEKEEPER

### Summary and Lesson Learned from These Experiments

Given the proof-of-concept evaluation of these results, the proposed methodology demonstrates its relevance for knowledge discovery research. One of the most interesting findings observed from the conducted experiments was the presence of some resistance and susceptibility genes among the candidates proposed by the method; these genes are known to play an important role in the plant defense process. Moreover, the best ranked candidate obtained for the *c* term inputs CCA1 and LHY (two central genes of the circadian clock rhythm) was DMR1 (a susceptibility gene, mutation of this gene results in a higher resistance), that is a hot topic of a plant resistance research lately. In future work, the genes identified in results will be closely inspected by domain experts. In conclusion, let us summarize this section by the lesson learned from these experiments.

#### Lesson Learned 6: Term

filtering and synonyms matter

In the experiments using plant defence-circadian

**Table 3** Evaluation according to rank

Rank	No, not at all	Not really	Maybe	Yes	P@15 Maybe	P@15 Yes
1.	2	4	1	8	0.067	0.533
2.	2	6	1	6	0.067	0.400
3.	1	7	1	6	0.067	0.400
4.	3	7	2	3	0.133	0.200
5.	1	3	2	9	0.133	0.600
6.	4	1	3	7	0.200	0.467
7.	1	4	4	6	0.267	0.400
8.	2	3	5	5	0.333	0.333
9.	1	6	3	5	0.200	0.333
10.	1	7	2	5	0.133	0.333
All	18	48	24	60	0.160	0.400

**Table 4** Evaluation on all output terms (duplicates removed)

Label	Count	Perc. (%)
NO, NOT AT ALL	14	19.18
NOT REALLY	19	26.03
MAYBE	13	17.81
YES	27	36.99
Total	73	100

rhythm domain pair, the goal was to identify potentially interesting new daily regulated mechanisms that are responsible for plant defence. After obtaining 5412 documents from PubMed containing complete articles (2483 from plant defence and 2929 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 the previous research projects (22,265 gene names mapped into 7863 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 6083 gene names), which resulted in a substantial reduction of the input file size (from 200 to 28 MB).

## Conclusions and Further Work

This paper addresses the field of scientific computational creativity, in particular bisociative literature-based discovery. The paper mostly focused 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, proposing a novel methodology exploiting the use of advanced embedding technology for bisociative cross-domain literature mining.

Our future work, aimed at improving the effectiveness of bridging term detection in cross-domain literature mining, will be performed in several directions, based on our current research: using ontologies for term enrichment in cross-domain document exploration, and using network analysis for cross-domain heterogeneous information network exploration.

- 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 the text representation learning to improve the performance and robustness of the learned models. Consequently, our novel `tax2vec` algorithm [34] could be used for constructing taxonomy-based features to improve the results of document clustering and classification.
- 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).
- We will also introduce additional user-interface options for data visualization and exploration, as well as advance our bridging term ranking methodology [17] by adding new heuristics, which will take into account also the semantic aspects of the data.
- Most importantly, we will further explore embeddings-based LBD in the closed LBD settings, aiming to improve and further explore the methodology proposed in “[Towards creative embeddings-based bisociative LBD](#)”. Especially, we plan to focus on bisociative discovery without known concept *c*, as well as on enabling multi-word expressions as output.
- We will experiment with new application topics. It will be especially insightful to address problems in need of discovering novel bisociations between two

different domains. Also, it could be useful to investigate two entirely unrelated domains to provide a baseline.

**Author Contributions** All authors contributed to the study conception and design. The first draft of the manuscript was written by Nada Lavrač, all authors contributed to manuscript writing, data collection, and analysis were performed by Matej Martinc, Senja Pollak, Maruša Pompe Novak, and Bojan Cestnik. All authors approved the final manuscript.

**Funding** This work was supported by the Slovenian Research Agency (ARRS) grants Knowledge technologies P2-0103, and Terminology and knowledge frames across languages (J6-9372), and the European Union's Horizon 2020 research and innovation programme under grant agreement No 825153, project EMBEDDIA (Cross-Lingual Embeddings for Less-Represented Languages in European News Media).

**Availability of data and material (data transparency)** Available upon request from the second author Matej Martinc.

## Compliance with Ethical Standards

**Conflict of interest** The authors declare that they have no conflict of interest.

**Code availability (software application or custom code)** Open source.

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

## References

1. Abgaz, Y., O'Donoghue, D., Hurley, D., Smorodinnikov, D.: Evaluation of analogical inferences formed from automatically generated representations of scientific publications. In: 24th Irish Conference on Artificial Intelligence and Cognitive Science (2016)
2. Agrawal, R., Mannila, H., Srikant, R., Toivonen, H., Verkamo, A.I., et al.: Fast discovery of association rules. *Adv. Knowl. Discov. Data Min.* **12**(1), 307–328 (1996)
3. Berthold, M. (ed.): *Bisociative Knowledge Discovery*. Springer, Berlin (2012)
4. Bojanowski, P., Grave, E., Joulin, A., Mikolov, T.: Enriching word vectors with subword information. *Trans. Assoc. Comput. Linguist.* **5**, 135–146 (2017)
5. Brodley, C.E., Friedl, M.A.: Identifying mislabeled training data. *J. Artif. Intell. Res.* **11**, 131–167 (1999)
6. Bruza, P., Weeber, M.: *Literature-based Discovery*. Springer Science & Business Media, Berlin (2008)

7. Cestnik, B., Fabbretti, E., Gubiani, D., Urbančič, T., Lavrač, N.: 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. *Genom. Comput. Biol.* **3**(3), e58 (2017)
8. Conneau, A., Lample, G., Ranzato, M., Denoyer, L., Jégou, H.: Word translation without parallel data. *CoRR abs/1710.04087* (2017)
9. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. (2018). *arXiv preprint arXiv:1810.04805*
10. Dong, F., O'Donoghue, D., Ersotelos, N., Wu, S., Saggion, H., Ronzano, F., Corcho, Ó., Hurley, D., Abgaz, Y.M., Zhang, J., Chaudhry, E., Yang, X., Wei, H., Deng, Z., Mahdian, B., Careil, J.M.: Dr. inventor, promoting scientific creativity by utilising web-based research objects. *Impact* **2**, 40–44 (2017)
11. 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: An Introduction to Concept, Algorithms, Tools, and Applications*, pp. 11–32. Springer, Berlin (2012)
12. Fortuna, B., Grobelnik, M., Mladenčić, D.: Semi-automatic data-driven ontology construction system. In: *Proceedings of the 9th International Multi-conference Information Society*, pp. 223–226 (2006)
13. Gopalakrishnan, V., Jha, K., Jin, W., Zhang, A.: A survey on literature based discovery approaches in biomedical domain. *J. Biomed. Inform.* **93**, 103141 (2019)
14. Henry, S., Cuffy, C., McInnes, B.T.: Vector representations of multi-word terms for semantic relatedness. *J. Biomed. Inform.* **77**, 111–119 (2018)
15. Holzinger, A., Yildirim, P., Geier, M., Simonic, K.M.: Quality-based knowledge discovery from medical text on the web. In: *Quality Issues in the Management of Web Information*, pp. 11–13. Springer (2013)
16. 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), 289–298 (2005)
17. Juršič, M., Cestnik, B., Urbančič, T., Lavrač, N.: Cross-domain literature mining: Finding bridging concepts with CrossBee. In: *Proceedings of the 3rd International Conference on Computational Creativity*, pp. 33–40 (2012)
18. Kastrin, A., Rindfleisch, T.C., Hristovski, D.: Link prediction on the semantic MEDLINE network. In: *Proceedings of the International Conference on Discovery Science*, pp. 135–143. Springer (2014)
19. Koestler, A.: *The Act of Creation*. Hutchinson, Paris (1964)
20. Kötter, T., Berthold, M.: From information networks to bisociative information networks. In: *Bisociative Knowledge Discovery*, pp. 33–50. Springer (2012)
21. Lavrač, N., Juršič, M., Sluban, B., Perovšek, M., Pollak, S., Urbančič, T., Cestnik, B.: Bisociative knowledge discovery for cross-domain literature mining. In: Veale, T., Cardoso, F.A. (eds.) *Computational Creativity: The Philosophy and Engineering of Autonomously Creative Systems*, pp. 121–139. Springer, Berlin (2019)
22. Lavrač, N., Martinc, M., Pollak, S., Cestnik, B.: Bisociative literature-based discovery: Lessons learned and new prospects. In: *Proceedings of International Conference on Computational Creativity (In press)* (2020)
23. Lindsay, R.K., Gordon, M.D.: Literature-based discovery by lexical statistics. *J. Am. Soc. Inf. Sci. Technol.* **1**, 574–587 (1999)
24. Martinc, M., Škrlić, B., Pirkmajer, S., Lavrač, N., Cestnik, B., Marzidovšek, M., Pollak, S.: Covid-19 therapy target discovery with context-aware literature mining. In: *Proceedings of International Conference on Discovery Science*. Springer (2020) (**In press**)
25. Mikolov, T., Chen, K., Corrado, G.S., Dean, J.: Efficient estimation of word representations in vector space. In: *Proceedings of ICLR CoRR*. abs/1301.3781 (2013)
26. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Burges, C.J.C., Bottou, L., Welling, M., Ghahramani, Z., Weinberger, K.Q. (eds.) *Advances in Neural Information Processing Systems*, vol. 26, pp. 3111–3119. Curran Associates Inc., New York (2013)
27. Mikolov, T., Yih, W.T., Zweig, G.: 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*, pp. 746–751. Association for Computational Linguistics (2013)

28. Pennington, J., Socher, R., Manning, C.: Glove: Global vectors for word representation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1532–1543. Association for Computational Linguistics, Doha, Qatar (2014)
29. Perovšek, M., Kranjc, J., Erjavec, T., Cestnik, B., Lavrač, N.: TextFlows: a visual programming platform for text mining and natural language processing. *Sci. Comput. Program.* **121**, 128–152 (2016)
30. Peters, M.E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., Zettlemoyer, L.: Deep contextualized word representations. (2018). arXiv preprint [arXiv:1802.05365](https://arxiv.org/abs/1802.05365)
31. Petrič, I., Cestnik, B., Lavrač, N., Urbančič, T.: Outlier detection in cross-context link discovery for creative literature mining. *Comput. J.* **55**(1), 47–61 (2012)
32. 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)
33. Sebastian, Y., Siew, E.G., Orimaye, S.O.: Emerging approaches in literature-based discovery: techniques and performance review. *Knowl. Eng. Rev.* **32**, e12 (2017)
34. Škrli, B., Martinc, M., Kralj, J., Lavrač, N., Pollak, S.: tax2vec: constructing interpretable features from taxonomies for short text classification. *Comput. Speech Lang.* **65**, 101104 (2021)
35. Sluban, B., Gamberger, D., Lavrač, N.: Ensemble-based noise detection: Noise ranking and visual performance evaluation. *Data Min. Knowl. Discov.* **28**, 1–39 (2013)
36. Sluban, B., Juršič, M., Cestnik, B., Lavrač, N.: Exploring the power of outliers for cross-domain literature mining. In: *Bisociative Knowledge Discovery*, pp. 325–337. Springer (2012)
37. Smalheiser, N., Swanson, D.R.: Using ARROWSMITH: a computer-assisted approach to formulating and assessing scientific hypotheses. *Comput. Methods Programs Biomed.* **57**(3), 149–154 (1998)
38. Srinivasan, P.: Text mining: generating hypotheses from MEDLINE. *J. Am. Soc. Inf. Sci. Technol.* **55**(5), 396–413 (2004)
39. Swanson, D.R.: Migraine and magnesium: eleven neglected connections. *Perspect. Biol. Med.* **78**(1), 526–557 (1988)
40. Swanson, D.R.: Medical literature as a potential source of new knowledge. *Bull. Med. Lib. Assoc.* **78**(1), 29 (1990)
41. 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. Technol.* **57**(11), 1427–1439 (2006)
42. Ulčar, M., Robnik-Šikonja, M.: Multilingual culture-independent word analogy datasets. In: Proceedings of LREC (2020) (**In press**)
43. Weber, M., Klein, H., de Jong-van den Berg, L., Vos, R., et al.: Using concepts in literature-based discovery: simulating Swanson's Raynaud-fish oil and migraine-magnesium discoveries. *J. Am. Soc. Inf. Sci. Technol.* **52**(7), 548–557 (2001)
44. Yetisgen-Yildiz, M., Pratt, W.: Using statistical and knowledge-based approaches for literature-based discovery. *J. Biomed. Inform.* **39**(6), 600–611 (2006)

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.