Chapter 6
Bisociative Knowledge Discovery for Cross-domain Literature Mining

Nada Lavrač, Matjaž Juršič, Borut Sluban, Matic Perovšek, Senja Pollak, Tanja Urbančič, and Bojan Cestnik

Abstract Given its immense growth, the scientific literature can be explored to reveal new discoveries, based on as yet undiscovered relations between knowledge from different, relatively isolated fields of specialization. This chapter presents an approach to creative knowledge discovery through the mechanism of *bisociation*. Bisociative reasoning is at the heart of creative, accidental discovery, i.e., serendipity. Bisociative knowledge discovery is focused on finding unexpected links by crossing between different contexts. In this work, bisociative knowledge discovery is explored in the framework of text mining, addressing cross-domain literature-based discovery. Two approaches are briefly outlined: the CrossBee approach to cross-domain bridging-term detection, and the OntoGen approach to bridging-term detection through outlier document exploration.

Nada Lavrač
Jožef Stefan Institute, Ljubljana, Jožef Stefan International Postgraduate School, Ljubljana and University of Nova Gorica, Slovenia. e-mail: nada.lavrac@ijs.si

Matjaž Juršič, Borut Sluban, Matic Perovšek
Jožef Stefan Institute, Ljubljana and Jožef Stefan International Postgraduate School, Ljubljana, Slovenia. e-mail: matjaz.jursic@ijs.si,borut.sluban@ijs.si,matic.perovsek@ijs.si

Senja Pollak
Jožef Stefan Institute, Ljubljana, Slovenia. e-mail: senja.pollak@ijs.si

Tanja Urbančič
University of Nova Gorica and Jožef Stefan Institute, Ljubljana, Slovenia.
e-mail: tanja.urbancic@ung.si

Bojan Cestnik
Temida d.o.o., Ljubljana and Jožef Stefan Institute, Ljubljana, Slovenia.
e-mail: bojan.cestnik@temida.si

T. Veale, F. A. Cardoso (eds.), *Computational Creativity*, Computational Synthesis and Creative Systems, https://doi.org/10.1007/978-3-319-43610-4_6

6.1 Introduction

The growing amounts of available knowledge and data exceed human analytic capabilities. Therefore new technologies that can help in 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 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, and in particular association rule learning (Agrawal, Mannila, Srikant, Toivonen, Verkamo, et al., 1996), scientific discovery requires creative thinking to connect seemingly unrelated information, for example, by using metaphors or analogies between concepts from different domains. These kinds of context-crossing associations, called *bisociations* (Koestler, 1964), are often needed for innovative discoveries.

This chapter provides an introduction to bisociative knowledge discovery, and outlines selected approaches to cross-domain literature mining that support experts in searching for hidden links connecting two seemingly unrelated domains, most notably the CrossBee approach to cross-domain bridging-term (b-term) detection (Juršič, Cestnik, Urbančič, & Lavrač, 2012a, 2012b), and the approach to cross-domain literature mining via outlier document detection and exploration (Petrič, Cestnik, Lavrač, & Urbančič, 2012; Sluban, Juršič, Cestnik, & Lavrač, 2012).

This chapter is organized as follows. Section 6.2 presents related work in the area of bisociative knowledge discovery, literature-based discovery (LBD) and the human–computer interaction (HCI) aspects of creativity support tools. Section 6.3 illustrates the problem of b-term ranking and exploration through a use case scenario, followed by an overview of the b-term detection and exploration methodology as implemented in the CrossBee exploration tool, including the ensemble heuristic used in b-term detection. Section 6.4 presents two approaches to outlier document detection that can be used to narrow down the search space of b-terms, given the fact that outlier documents contain most of the cross-domain b-terms, as shown in our past research. Finally, Section 6.5 concludes with a summary of the methods presented and directions for further work.

6.2 Related Work

This section presents related work in the area of bisociative knowledge discovery, LBD and the HCI aspects of creativity support tools.

6.2.1 Bisociative Knowledge Discovery

Bisociative knowledge discovery is a challenging task motivated by a trend of over-specialization in 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. Information that is related across different contexts is difficult to identify using associative approaches, like standard association rule learning (Agrawal et al., 1996) known from the data-mining and machine learning literature. Therefore, the ability of literature-mining methods and software tools to support experts in their knowledge discovery processes – especially in searching for yet unexplored connections between different domains – is becoming increasingly important.

Koestler (1964) 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 the arts. As an example of bisociative scientific discovery, Koestler (p. 105) cites the "Eureka" discovery of Archimedes, bisociating the measurement of the volume of nonregular solids with the displacement of water:

> No doubt he had observed many times that the level of the [bath] water rose whenever he got into it; but this fact, and the distance between the two levels, was totally irrelevant to him – until it suddenly became bisociated with his problem. At that instant he realised that the amount of rise of the water-level was a simple measure of the volume of his own complicated body.

The concept of bisociation is illustrated in Fig. 6.1. It should be noted that context crossing is subjective, since the user has to move from their "normal"' context (frame of reference) to a *habitually incompatible context* to find the bisociative link. In Koestler's terms (Fig. 6.1), 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). The creative act here is to find links (from S to the target T) which lead "out-of-the-plane" via intermediate, bridging concepts (L). Thus, contextualization and link discovery are two of the fundamental mechanisms in bisociative reasoning.

In summary, according to Koestler (1964), 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 esthetic experience. Koestler regarded many different mental phenomena

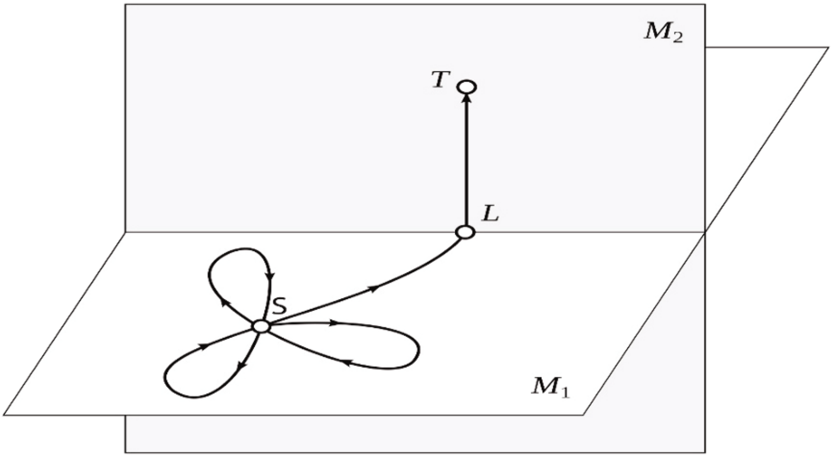


Fig. 6.1 Koestler's schema of bisociative discovery in science (Koestler, 1964, p. 107).

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 researchers interested in so-called *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. 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 modern terms (Berthold, 2012), bisociative knowledge discovery thus addresses a data-mining task where two or more domains of interest are searched for bisociative links or bridging concepts (i.e., individual context-bridging terms). Note that in this context, a single *domain* does not necessary refer to a single feature space; instead, we use this term to denote that the objects under analysis all represent properties with respect to one – more or less specific – aspect, even with multiple representations of the same space of objects (multiview learning, parallel universes, and redescription mining are well-known techniques addressing multiple representations of objects in the same domain of discourse). In contrast, bisociative knowledge discovery looks for connections between different domains of discourse, using either the same representation of different domains or different domain representations, where – according to Berthold (2012) – bridging concepts can be detected as nodes bridging different graphs, as subgraphs linking different graphs, as bridging links in terms of graph similarity, or as bridging terms appearing in different document corpora. The latter, referred to as *bridging-term discovery*, is the focus of the research described in this chapter.

6.2.2 Literature-Based Discovery

In LBD (Bruza & Weeber, 2008) – and, in particular, in cross-domain literature mining, which addresses knowledge discovery in two (or several) initially separate document corpora – a crucial step is the identification of interesting b-terms that carry the potential tp revealing the links connecting the separate domains. As shown by Petrič et al. (2012), LBD (Bruza & Weeber, 2008) is closely related to bisociative knowledge discovery (Berthold, 2012); for example, the b-terms known from the LBD literature directly correspond to Koestler's notion of bridging concepts L, introduced in the previous section.

The early work in LBD was due to Swanson (1990) and Smalheiser and Swanson (1998), who developed an approach to assist a user in LBD by detecting interesting cross-domain terms with the goal of uncovering the possible relations between previously unrelated concepts. 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 b-terms are investigated further by the user for their potential to generate new scientific hypotheses.[1] Their approach, known as the "ABC model of knowledge discovery", addresses several settings, including the *closed discovery* setting (Weeber, Klein, de Jong-van den Berg, Vos, 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 a bridging concept (term) b in B in order 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. 6.2.

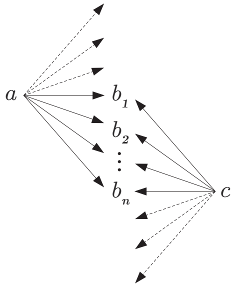


Fig. 6.2 Closed discovery process defined by Weeber, Klein, de Jong-van den Berg, Vos, et al. (2001).

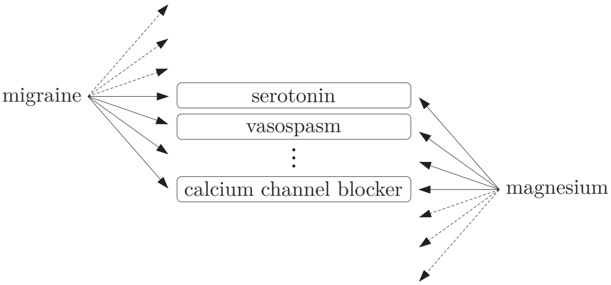


Fig. 6.3 Closed discovery when exploring migraine and magnesium documents, with b-terms identified by Swanson, Smalheiser, and Torvik (2006).

[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.

Swanson's seminal work showed that databases such as PubMed can serve as a rich source of hidden relations between usually unrelated topics, potentially leading to novel insights and discoveries. By studying two separate literatures – the literature on migraine headache and the articles on magnesium – Swanson (1988) discovered "Eleven neglected connections", all of them supportive of the hypothesis that magnesium deficiency might cause migraine headache. Figure 6.3 illustrates the closed discovery setting for Swanson's task of finding the terms linking the "migraine" and "magnesium" domains. Swanson's literature-mining results were later confirmed by laboratory and clinical investigations. This well-known example has become the gold standard in the literature-mining field and has been used as a benchmark in several studies (Juršič et al., 2012b; Lindsay & Gordon, 1999; Srinivasan, 2004; Weeber et al., 2001).

Inspired by this early work, literature-mining approaches were developed further and successfully applied to different problems, such as finding associations between genes and diseases (Hristovski, Peterlin, Mitchell, & Humphrey, 2005), between diseases and chemicals (Yetisgen-Yildiz & Pratt, 2006), and others. Holzinger, Yildirim, Geier, and Simonic (2013) described several quality-oriented web-based tools for the analysis of the biomedical literature, which include analysis of terms (biomedical entities such as diseases, drugs, genes, proteins, and organs) and provide concepts associated with a given term. A more recent approach by Kastrin, Rindflesch, and Hristovski (2014) is complementary to the other LBD approaches, as it uses different similarity measures (such as common neighbors, the Jaccard index, and preferential attachment) for link prediction of implicit relationships in the Semantic MEDLINE network.

Supporting the user in effectively searching for b-terms provided a motivation for developing the CrossBee approach to b-term detection applicable in the closed discovery setting (Juršič et al., 2012b), implemented through ensemble-based term ranking, where an ensemble heuristic composed of six elementary heuristics was constructed for term evaluation. This approach is described in more detail in Section 6.3. Furthermore, the research conducted by Petrič et al. (2012) and Sluban et al. (2012) suggests that b-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. (2012) have shown that such documents, considered as outlier documents of their own domain, contain a substantially larger amount of bridging/linking terms than the regular nonoutlier documents. This approach, using the OntoGen tool (Fortuna, Grobelnik, & Mladenić, 2006), is described in more detail in Section 6.4.

In conclusion, let us summarize the relationship between bisociative knowledge discovery and Swanson's ABC model of 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. Petrič et al. (2012) have presented a unifying view that establishes relationships between the two frameworks, as summarized in Table 6.1. Similarly to a bisociation, which, according to Koestler, is a result of processes of the mind when making new associations between concepts S and T from usually separated contexts (illustrated in Fig. 6.1), literature-based discoveries in Swanson's ABC model are a result of uncovering links between con-

Table 6.1 Unifying Koestler's and Swanson's models of creative knowledge discovery (Petrič, Cestnik, Lavrač, & Urbančič, 2012)

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 $b \in \text{terms}(A) \cap \text{terms}(C)$

cepts a and c from disjoint literatures A and C (illustrated in Fig. 6.2). In terms of Koestler's model, the two domains A and C, investigated in the closed LBD framework, correspond to the two habitually incompatible frames of reference, M_1 and M_2. Moreover, the b-terms $b_1, b_2, \ldots, b_n$ that are common to literatures 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.

6.2.3 *Creativity Support Tools and HCI*

CrossBee and the outlier detection tools developed can be viewed as creativity support tools, which are closely related to the field of HCI, as stated by Resnick et al. (2005) when summarizing the aims of designing creativity support tools (CSTs) as follows:

> Our goal is to develop improved software and user interfaces that empower users to be not only more productive, but more innovative.

The work of Shneiderman (2007, 2009) provides a structured set of design principles for CSTs, outlined below:

- *Support exploration*. To be successful in discovery and innovation, users should have access to improved search services providing rich mechanisms for organizing search results by ranking, clustering, and partitioning, with ample tools for annotation, tagging, and marking.
- *Enable collaboration*. While the actual discovery moments in innovation can be very personal, the processes that lead to them are often highly collaborative.
- *Provide rich history-keeping*. The benefits of rich history-keeping are that users have a record of which alternatives they have tried, they can compare the many alternatives, and they can go back to earlier alternatives to make modifications.
- *Design with low thresholds, high ceilings, and wide walls*. CSTs should have a short learning curve for novices (low threshold), yet provide sophisticated functionality that experts need (high ceilings), and also deliver a wide range of supplementary services to choose from (wide walls).

These principles were followed in our implementations and used in the evaluation of our approaches outlined in Sections 6.3 and 6.4 below, using two creativity support tools CrossBee (Juršič et al., 2012a, 2012b) and OntoGen (Fortuna et al., 2006), respectively:

- *CrossBee* (Juršič et al., 2012a, 2012b) is an off-the-shelf solution for finding bisociations bridging two user-defined domains (separate domain literatures). CrossBee is a system that suggests b-terms using an ensemble-based term-ranking methodology. The tool also helps experts in searching for hidden links that connect two seemingly unrelated domains. In addition to this core functionality, supplementary functionality and content presentations have been added, which make the CrossBee web application a user-friendly tool for the ranking and exploration of prospective cross-context links. This enables the user not only to spot but also to efficiently investigate the cross-domain links discovered. The CrossBee user-friendly human–computer interface is briefly presented in Section 6.3.4.
- *OntoGen* (Fortuna et al., 2006) is a semiautomatic data-driven interactive text-mining tool that aids the user during the creative process of topic ontology construction. In essence, it is a text-mining tool for grouping documents into cohesive clusters, which can be considered as concepts in an automatically constructed topic ontology. The underlying methodology is k-means clustering, which is a particularly popular technique, since only the parameter k needs to be chosen to determine the number of categories in to which the documents will be clustered. A particularly appealing feature is OntoGen's user-friendly human–computer interface. The "main window" provides ontology visualization, where each concept is represented by the top three keywords (automatically assigned names of clusters, which can be manually edited), while the "concept hierarchy" window offers a quick overview of all the concepts with their positions in the concept hierarchy, which can also be directly manipulated. A particular use of OntoGen for outlier document detection is described in Section 6.4.2.

6.3 Bridging-Term Detection in Literature-Based Discovery

This section briefly describes our previous work on bisociative knowledge discovery in the area of literature mining, focusing on the CrossBee methodology for b-term detection, outlined in Juršič et al. (2012a, 2012b).

6.3.1 CrossBee Methodology

In cross-domain knowledge discovery, estimating which of the terms have a high potential for interesting discoveries is a challenging research question. It is especially important for cross-context scientific discovery such as understanding complex medical phenomena or finding new drugs for illnesses yet not fully understood.

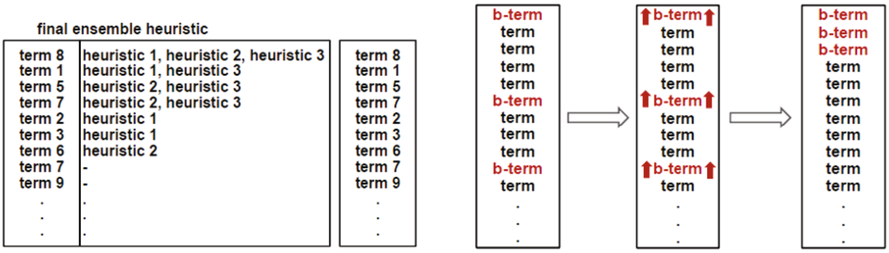
The ensemble-based ranking methodology for b-term detection is illustrated in Fig. 6.4, showing the term ranking using an ensemble heuristic. Figure 6.5 shows the list of b-terms ranked by voting of an ensemble of heuristics, where the ranked list presented is the actual output produced by the CrossBee b-term exploration system using the gold standard dataset in literature mining, i.e., the combined migraine–magnesium dataset (Swanson, 1988). The ranked list of b-term candidates, shown in Figure 6.5, provides the user with some additional information, including the votes of the individual base heuristics in the ensemble and the domain occurrence statistics of the terms in both domains.

6.3.2 *Heuristics for Bridging-Term Discovery*

Several different elementary and ensemble heuristics for b-term ranking are available in CrossBee. The heuristics are defined as functions that numerically evaluate the quality of a term by assigning a bisociation score to it (measuring the potential that a term is actually a b-term). For the definition of an appropriate set of heuristics, we define a set of special (mainly statistical) properties of terms, which are aimed at distinguishing b-terms from regular terms; thus, these heuristics can also be viewed as advanced term statistics.

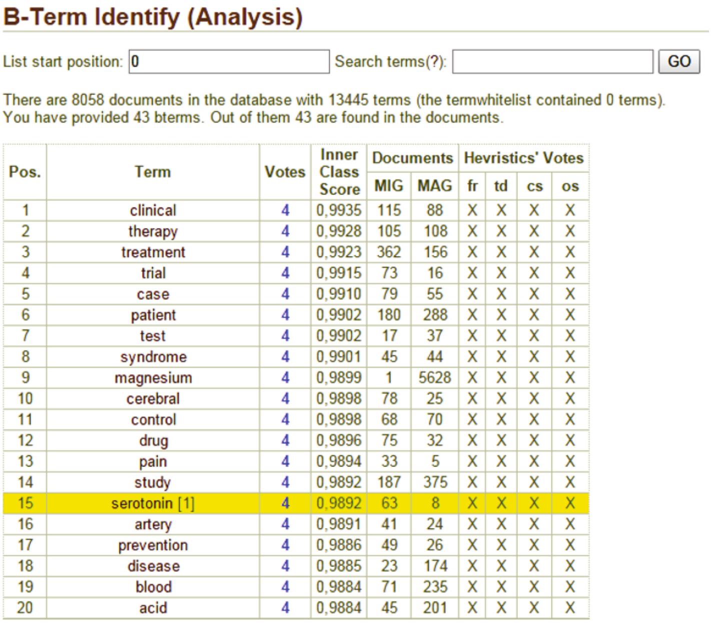
Formally, a heuristic is a function with two inputs, i.e., a set of domain-labeled documents D and a term t appearing in those documents, and one output, i.e, a score that represents the term's bisociation potential. All of the heuristics operate on data retrieved from the documents in text preprocessing. While term ranking using scores calculated by an ideal heuristic should result in ranking all the b-terms at the top of the ranked list, this ideal scenario is not realistic; nevertheless, ranking by heuristic scores should at least increase the proportion of b-terms at the top of the ranked term list.

We use the following notation: to state that a term's bisociation score b_score is equal to the result of a heuristic named $heurX$, we can write $b\_score = heurX(D,t)$. However, since the set of input documents is static when we are dealing with a


final ensemble heuristic

term 8	heuristic 1, heuristic 2, heuristic 3	term 8
term 1	heuristic 1, heuristic 3	term 1
term 5	heuristic 2, heuristic 3	term 5
term 7	heuristic 2, heuristic 3	term 7
term 2	heuristic 1	term 2
term 3	heuristic 1	term 3
term 6	heuristic 2	term 6
term 7	-	term 7
term 9	-	term 9
.	.	.
.	.	.
.	.	.

Fig. 6.4 Term-ranking approach: first, ensemble heuristics vote for terms, and next, terms are sorted according to their b-term potential (as shown on the left). Consequently, b-terms with the highest b-term potential should receive the highest scores (as shown on the right).



B-Term Identify (Analysis)

There are 8058 documents in the database with 13445 terms (the termwhitelist contained 0 terms).
You have provided 43 bterms. Out of them 43 are found in the documents.

Pos.	Term	Votes	Inner Class Score	Documents		Hevristics' Votes			
				MIG	MAG	fr	td	cs	os
1	clinical	4	0,9935	115	88	X	X	X	X
2	therapy	4	0,9928	105	108	X	X	X	X
3	treatment	4	0,9923	362	156	X	X	X	X
4	trial	4	0,9915	73	16	X	X	X	X
5	case	4	0,9910	79	55	X	X	X	X
6	patient	4	0,9902	180	288	X	X	X	X
7	test	4	0,9902	17	37	X	X	X	X
8	syndrome	4	0,9901	45	44	X	X	X	X
9	magnesium	4	0,9899	1	5628	X	X	X	X
10	cerebral	4	0,9898	78	25	X	X	X	X
11	control	4	0,9898	68	70	X	X	X	X
12	drug	4	0,9896	75	32	X	X	X	X
13	pain	4	0,9894	33	5	X	X	X	X
14	study	4	0,9892	187	375	X	X	X	X
15	serotonin [1]	4	0,9892	63	8	X	X	X	X
16	artery	4	0,9891	41	24	X	X	X	X
17	prevention	4	0,9886	49	26	X	X	X	X
18	disease	4	0,9885	23	174	X	X	X	X
19	blood	4	0,9884	71	235	X	X	X	X
20	acid	4	0,9884	45	201	X	X	X	X

Fig. 6.5 The ensemble-heuristic-based ranking page, indicating with a cross (X) which elementary heuristics have identified the term as a potential b-term. This example shows the 20 top-ranked terms from the migraine–magnesium domain according to the selected heuristics.

concrete dataset, we can – for the sake of simplicity – omit the set of input documents from the notation for the heuristic and use $b\_score = heurX(t)$. Whenever we need to explicitly specify the set of documents to which a function is applied (this is never needed for a heuristic, but sometimes needed for auxiliary functions used in the formula for the heuristic), we write it as $funcX_D(t)$. To specify the function's input document set, we have two options: we can either use D_u, which stands for the (union) set of all the documents from all the domains, or use $D_n : n \in \{1..N\}$, which stands for the set of documents from the given domain n. In general, the following statement holds: $D_u = \cup_{n=1}^{N} D_n$, where N is the number of domains. In the most common scenario, when there are exactly two distinct domains, we also use the notation D_A for D_1 and D_C for D_2, similarly to Swanson's notation using the symbols A and C as representatives of the initial and the target domain in the closed discovery setting, as mentioned in Section 6.2.

We defined four sets of base heuristics: six frequency-based, four TF-IDF-weight-based ("TF-IDF" denotes the product of term frequency and inverse document frequency weights, frequently used in document vector representations in text mining (Salton & Buckley, 1988)), three similarity-based, and eight outlier-based heuristics. Most of the heuristics that we developed work in a fundamentally similar way – they all manipulate solely the data present in the term and document vectors and derive the bisociation score of the terms. The exceptions to this are the outlier-based heuristics, which first evaluate outlier documents and only later use the information from the term vectors for b-term evaluation. Using these base heurisctics, we developed the ensemble heuristic described below.

6.3.3 Ensemble Heuristic

The ensemble heuristic for b-term discovery, which we constructed based on the experiments, is constructed as a sum of two parts, $s_t = s_t^{\text{vote}} + s_t^{\text{pos}}$, i.e., the ensemble voting score s_t^{vote} and the ensemble position score s_t^{pos}, which are summed together to give the final ensemble score for every term in the corpus vocabulary. Each term score represents the term's potential for linking the two disjoint domains.

The ensemble voting score (s_t^{vote}) of a given term t is an integer, which denotes how many base heuristics voted for the term: each term can be given a score $s_{t_j}^{\text{vote}} \in \{0,1,2,\ldots,k\}$, where k is the number of base heuristics used in the ensemble. The ensemble voting score of term t_j at position p_j in the ranked list of n terms is computed as a sum of the voting scores of the individual heuristics:

$$s_{t_j}^{\text{vote}} = \sum_{i=1}^{k} s_{t_j,h_i}^{\text{vote}} = \sum_{i=1}^{k} \begin{cases} 1, & p_j < n/3, \\ 0, & \text{otherwise.} \end{cases} \tag{6.1}$$

The ensemble position score (s_t^{pos}) is calculated as an average of the position scores of the individual base heuristics. For each heuristic h_i, the term's position score s_{t_j,h_i}^{pos} is calculated as $n - p_j/n$, which results in the position scores being in the interval $[0,1)$. For an ensemble of k heuristics, the ensemble position score is computed as an average of the position scores of the individual heuristics:

$$s_{t_j}^{\text{pos}} = \frac{1}{k}\sum_{i=1}^{k} s_{t_j,h_i}^{\text{pos}} = \frac{1}{k}\sum_{i=1}^{k} \frac{n - p_j}{n}. \tag{6.2}$$

The method of constructing the ensemble score described above looks rather intricate; however, the calculation of the ensemble score by our method is well justified by extensive experimental results (Juršič et al., 2012a, 2012b) on the migraine–magnesium dataset (Swanson, 1988). Based on the experimental results, the final set of elementary heuristics included in the ensemble consisted of the following heuristics:

- outFreqRelRF, the relative frequency of term t in the outlier document set detected by a random forest classifier;
- outFreqRelSVM, the relative frequency of term t in the outlier document set detected by a support vector machine classifier;
- outFreqRelCS, the relative frequency of term t in the outlier document set detected by a centroid similarity classifier;
- outFreqSum, the sum of the frequencies of term t in all three outlier document sets;

- tfidfDomnSum, the sum of the TF-IDF weights of term t in the two domains; and
- freqRatio, the term-to-document frequency ratio.

A detailed justification for the choice of this particular combination of heuristics is presented in Juršič (2015).

6.3.4 The CrossBee HCI Interface

The user-friendly CrossBee web interface can be used to efficiently investigate cross-domain links ranked by the ensemble-based ranking methodology. CrossBee's document-focused exploration empowers the user to filter and order the documents by various criteria, including a detailed document view that provides a more detailed presentation of a single document, including various term statistics. Methodology performance analysis supports the evaluation of the methodology by providing various data which can be used to measure the quality of the results, for example data for plotting ROC curves. High-ranked-term emphasis marks the terms according to their bisociation score calculated by the ensemble heuristic. When this feature is used, all high-ranked terms are emphasized throughout the whole application, thus making them easier to spot (see the different font sizes in Fig. 6.6). B-term emphasis marks the terms defined as b-terms by the user (terms highlighted in yellow in Fig. 6.6). Domain separation is a simple but effective option which colors all documents from the same domain in the same color, making an obvious distinction between the documents from the two domains (different colors in Fig. 6.6). User interface customization enables the user to decrease or increase the intensity of the following features: high-ranked term emphasis, b-term emphasis, and domain separation.

The user can inspect the actual appearances of the selected term in both domains, using side-by-side document inspection as shown in Fig. 6.6. In this way, they can verify whether their rationale behind selecting this term as a b-term can be justified based on the contents of the documents inspected.

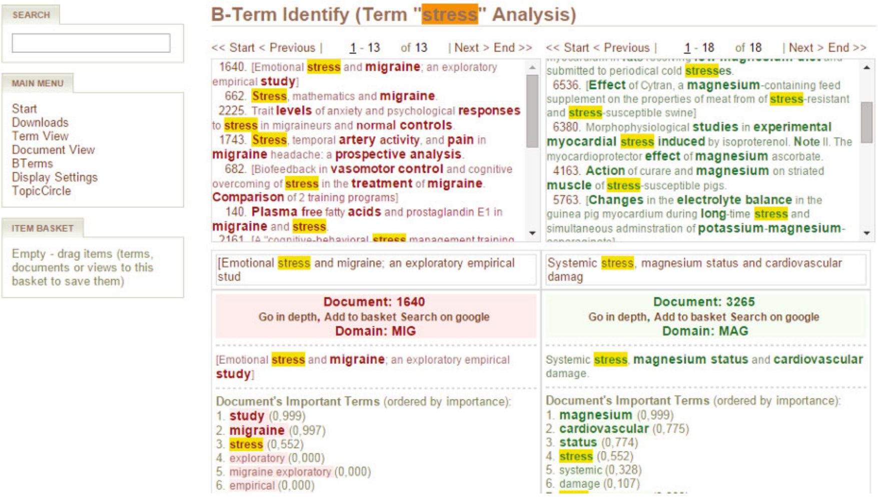
6.4 Exploring Outlier Documents in Literature-Based Discovery

This section outlines the exploration of outlier documents as means for cross-domain LBD (Petrič et al., 2012; Sluban et al., 2012).Here, we use the term "outlier detection" to refer to the task of finding irregular or unusual data instances (documents in the case of literature mining) that do not conform to the expected distribution.

Outlier detection is an established area of data mining (Aggarwal, 2013). Conceptually, an outlier is an unexpected event or entity, or – in our case – an irregular document. We are especially interested in outlier documents since they frequently embody new information that is hard to explain in the context of existing knowledge. Moreover, in data mining, an outlier is occasionally a primary object of study, as it can potentially lead to the discovery of new knowledge. These assumptions are well aligned with the bisociation potential that we wish to optimize; thus, we have constructed several heuristics that harvest the information possibly residing in outlier documents.

6.4.1 *Outlier Document Detection and B-term Identification Through Document Classification*

The technique proposed by Sluban et al. (2012) to detect outlier documents using classification algorithms, works as follows. Having documents from two domains of interest, we first train a classification model that distinguishes between the documents from these domains. Using the model constructed we classify all the documents. The documents that are misclassified – according to their domain of origin – are declared to be outlier documents, since according to the classification model they do not belong to their domain of origin. These domain outliers are actually borderline documents, as they were considered by the model to be more similar to the other domain than their originating domain. Hence they can be regarded as bridging documents between the two domains.


SEARCH
MAIN MENU
Start
Downloads
Term View
Document View
BTerms
Display Settings
TopicCircle
ITEM BASKET
Empty - drag items (terms, documents or views to this basket to save them)
B-Term Identify (Term "stress" Analysis)
<< Start < Previous | 1 - 13 of 13 | Next > End >>
<< Start < Previous | 1 - 18 of 18 | Next > End >>
Go in depth, Add to basket Search on google
Go in depth, Add to basket Search on google

1640. [Emotional stress and migraine; an exploratory empirical study]
662. Stress, mathematics and migraine.
2225. Trait levels of anxiety and psychological responses to stress in migraineurs and normal controls.
1743. Stress, temporal artery activity, and pain in migraine headache: a prospective analysis.
682. [Biofeedback in vasomotor control and cognitive overcoming of stress in the treatment of migraine. Comparison of 2 training programs]
140. Plasma free fatty acids and prostaglandin E1 in migraine and stress.

submitted to periodical cold stresses.
6536. [Effect of Cytran, a magnesium-containing feed supplement on the properties of meat from of stress-resistant and stress-susceptible swine]
6380. Morphophysiological studies in experimental myocardial stress induced by isoproterenol. Note II. The myocardioprotector effect of magnesium ascorbate.
4163. Action of curare and magnesium on striated muscle of stress-susceptible pigs.
5763. [Changes in the electrolyte balance in the guinea pig myocardium during long-time stress and simultaneous adminstration of potassium-magnesium-

[Emotional stress and migraine; an exploratory empirical stud

Systemic stress, magnesium status and cardiovascular damag

Document: 1640
Domain: MIG

[Emotional stress and migraine; an exploratory empirical study]

Document's Important Terms (ordered by importance):
1. study (0,999)
2. migraine (0,997)
3. stress (0,552)
4. exploratory (0,000)
5. migraine exploratory (0,000)
6. empirical (0,000)

Document: 3265
Domain: MAG

Systemic stress, magnesium status and cardiovascular damage.

Document's Important Terms (ordered by importance):
1. magnesium (0,999)
2. cardiovascular (0,775)
3. status (0,774)
4. stress (0,552)
5. systemic (0,328)
6. damage (0,107)

Fig. 6.6 One of the useful features of the CrossBee interface is the side-by-side view of documents from the two domains under investigation. The analysis of the b-term “stress” from the migraine–magnesium domain is shown. The view presented enables efficient comparison of two documents, the left one from the migraine domain and the right one from the magnesium domain.

In our work, we thus used noise detection approaches to find outlier documents containing cross-domain b-terms between two different domains. When exploring a domain pair dataset, we searched for a set of outlier documents using different classification-noise-filtering approaches (Brodley & Friedl, 1999), implemented and adapted for this purpose.

Classification noise filtering is based on the idea of using a classifier as a tool for detecting noisy and outlier instances in data. In this work, the simple classifiers used by Brodley and Friedl (1999) were replaced by new, better-performing classifiers, as the noise filter should, as much as possible, trust the classifiers that they will be able to correctly predict the class of a data instance. In this way, the incorrectly classified instances are considered to be noise/outliers. In other words, if an instance of class *A* is classified in the opposite class *C*, we consider it to be an outlier of domain *A*, and vice versa. We denote the two sets of domain outlier documents by $\mathscr{O}$(A) and $\mathscr{O}$(C), respectively. Figure 6.7 illustrates the principle.

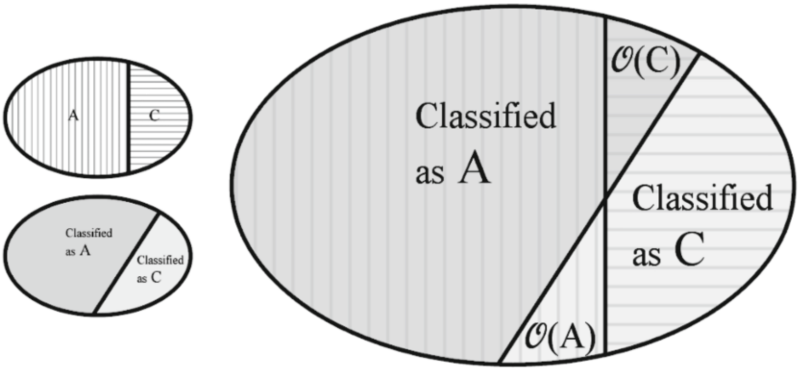


Fig. 6.7 Detecting outliers of a domain pair dataset using document classification.

We evaluated whether domain outliers obtained by classification noise filtering have the potential for bridging different concepts. We tested this on the migraine–magnesium (Swanson, Smalheiser, & Torvik, 2006) and autism–calcineurin (Petrič, Urbančič, Cestnik, & Macedoni-Lukšič, 2009) domain pair datasets, which have lists of confirmed concept b-terms. The experimental results showed that the sets of detected outlier documents were relatively small – including less than 5% of the entire datasets – and that they contained a great majority of b-terms; the number of b-terms in them was significantly higher than in same-sized random subsets. These results are summarized in Fig. 6.8. Hence the effort needed for finding cross-domain links is substantially reduced, as it requires one to explore a much smaller subset of documents, where a great majority of the b-terms are present and these terms are more frequent.

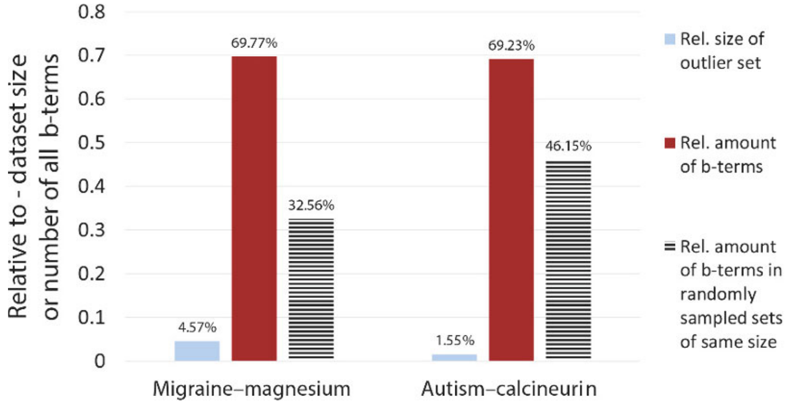


Fig. 6.8 Presence of b-terms in the detected outlier sets of two domain pair datasets.

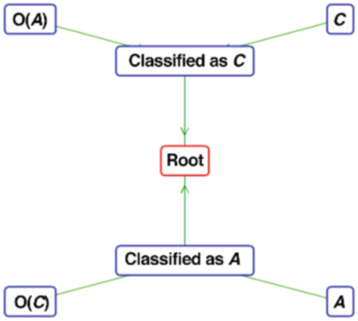


Fig. 6.9 Target domain documents from literatures A and C, clustered according to OntoGen's two-step approach, using first unsupervised and then supervised clustering to obtain outlier documents $\mathcal{O}(\mathrm{A})$ and $\mathcal{O}(\mathrm{C})$ of literatures A and C, respectively.

6.4.2 *Outlier Document Detection and B-term Identification Through Document Clustering*

The approach proposed by Petrič et al. (2012) concentrates on a specific type of outlier – the domain outliers – i.e., the documents that tend to be more similar to the documents in the opposite domain than to those in their own domain. In this approach, document clustering is used to find outlier documents. The approach consists of two steps. In the first step, the OntoGen clustering algorithm (Fortuna et al., 2006) is applied to cluster the merged document set $A \cup C$, consisting of documents from both of the domains A and C. The result of unsupervised clustering is two document clusters: A' = *Classified as A* (i.e., documents from $A \cup C$ classified as A), and C' = *Classified as C* (i.e., documents from $A \cup C$ classified as C). Then, in the second step, for each of the clusters, a supervised clustering approach is applied taking into account the original domains A and C of the documents. As a result, a two-level tree hierarchy of clusters is generated. The approach is illustrated in Fig. 6.9.

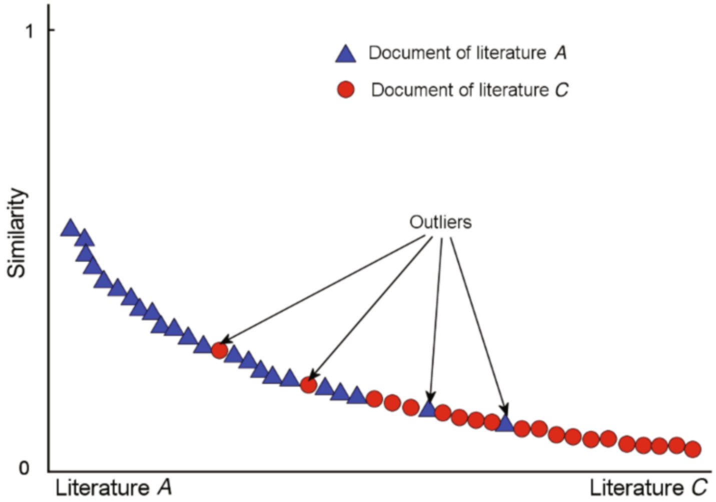


Fig. 6.10 Graph representing instances (documents) of literature *A* and instances (documents) of literature *C* according to their content similarity to a prototypical document of literature *A*, as suggested in Petrič, Cestnik, Lavrač, and Urbančič (2012). In this graph, outliers of literature *C* are positioned closer to the typical representatives of literature *A* than to the central documents of literature *C*.

The experimental results obtained in the gold standard migraine–magnesium domain, as well as in the autism–calcineurin domain pair, confirm the hypothesis that most b-terms appear in outlier documents and that, by considering only outlier documents, the search space for b-term identification can be greatly reduced. Moreover, the user can employ the document similarity graph – schematically presented in Fig. 6.10 – to identify the most irregular documents in their own domain and start the search for b-terms from these outlier documents, belonging to subclusters $\mathscr{O}(\mathrm{C})$ and $\mathscr{O}(\mathrm{A})$. In this way, the search space for finding b-term candidates can be substantially reduced.

6.4.3 Relating Outlier Document Detection to CrossBee Heuristics

The outlier document detection approaches described above inspired the development of outlier-based heuristics for the CrossBee b-term detection engine. As mentioned in Section 6.3.3, six heuristics (outFreqRelRF, outFreqRelSVM, outFreqRelCS, outFreqSum, tfidfDomnSum, and freqRatio) are used in the CrossBee ensemble heuristic. The outlier-based heuristics proved to be very effective. Note that four of these (outFreqRelRF, outFreqRelSVM, outFreqRelCS, and outFreqSum) are based on term frequencies in outlier documents; three of them were inspired by the classification-based approach (outFreqRelRF, outFreqRelSVM, and outFreqRelCS) and one by the OntoGen clustering approach (outFreqSum) to outlier document detection.

6.5 Conclusions and Further Work

This chapter has presented selected information technologies for creative knowledge discovery, developed to uncover previously unknown links between facts in different contexts, potentially leading to new insights and new knowledge. The approaches described are based on the Koestler's notion of bisociations, connecting domains that are usually considered as separate. When the domains investigated are described by texts, for example a set of documents, bisociative literature-mining methods can point towards novel chains of thoughts by identifying bridging terms with high potential for new discoveries resulting from putting existing pieces of knowledge together into a novel, interesting and reasonable whole.

The identification of cross-context links or bridging terms leading to new insights and discoveries is not an easy task, owing to the huge search space of possibilities, similar to looking for a needle in a haystack. One of the possible solutions is to identify the parts of the search space with an increased probability of finding good candidate terms/concepts with, the aim of restricting the huge amount of existing literature to a more manageable amount of sources to be explored first. This is the approach taken in our research in cross-domain literature mining via outlier document detection and exploration, presented in Section 6.4. The other option is to estimate the potential of candidate links for new discoveries and to concentrate on the most promising ones. This is the approach taken in the development of the online CrossBee application, supporting the user in the search and detection of cross-domain bridging terms, outlined in Section 6.3. The information technologies outlined, and other related approaches to bisociative link discovery that help in uncovering new connections between existing pieces of knowledge in the literature, can be used to assist researchers in their creative process by suggesting and even ranking candidate bridging terms.

Note that IT tools that implement literature-based discovery to enable a researcher to guide the discovery process by using his or her background knowledge enable the researcher to explore the literature more efficiently, but may also trigger the researcher's own human creativity. IT-supported literature exploration may provoke the researcher's own 'Koestler-style' bisociations to be triggered in this process. These bisociations may better specify or redirect the focus of further steps in the literature-based discovery process. The history of science and engineering offers numerous examples showing that Koestler's bisociative principle of thought has been an important element of new discoveries, based on innovative connections between already known ideas. As described above, literature-based discovery and bisociative knowledge discovery can complement each other, offering immense possibilities for new discoveries that we have only started to explore, but have already seen this process working at its best.

In future work we will 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. Besides, we will apply the system to new domain pairs to exhibit its generality, investigate the need and possibilities of dealing with domain specific background knowledge, and assist researchers in different disciplines in their explorations which may lead to new scientific discoveries. We will also propose a further extension of the literature based discovery methodology by facilitating the use of controlled vocabularies, enhancing the heuristics capability to rank the actual b-terms at the top of the ranked term list.

Acknowledgements This work was supported by the Slovenian Research Agency and the FP7 European Commission FET project PROSECCO (grant 600653).

References

Aggarwal, C. (2013). *Outlier analysis*. Springer.

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.

Berthold, M. (Ed.). (2012). *Bisociative knowledge discovery*. Springer.

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

Bruza, P., & Weeber, M. (2008). *Literature-based discovery*. Springer Science & Business Media.

Fortuna, B., Grobelnik, M., & Mladenić, D. (2006). Semi-automatic data-driven ontology construction system. In *Proceedings of the 9th International Multiconference Information Society* (pp. 223–226).

Holzinger, A., Yildirim, P., Geier, M., & Simonic, K.-M. (2013). Quality-based knowledge discovery from medical text on the web. In G. Pasi, G. Bordogna, & L. C. Jain (Eds.), *Quality issues in the management of web information* (pp. 11–13). Springer.

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

Juršič, M. (2015). *Text mining for cross-domain knowledge discovery* (Doctoral dissertation, Jožef Stefan International Postgraduate School, Ljubljana, Slovenia).

Juršič, M., Cestnik, B., Urbančič, T., & Lavrač, N. (2012a). Bisociative literature mining by ensemble heuristics. In M. Berthold (Ed.), *Bisociative knowledge discovery* (pp. 338–358). Springer.

Juršič, M., Cestnik, B., Urbančič, T., & Lavrač, N. (2012b). Cross-domain literature mining: Finding bridging concepts with CrossBee. In *Proceedings of the 3rd international conference on computational creativity* (pp. 33–40).

Kastrin, A., Rindflesch, T. C., & Hristovski, D. (2014). Link prediction on the semantic MEDLINE network. In S. Džeroski, P. Panov, D. Kocev, & L. Todorovski (Eds.), *Discovery science* (pp. 135–143). Springer.

Koestler, A. (1964). *The act of creation*. Hutchinson.

Lindsay, R. K., & Gordon, M. D. (1999). Literature-based discovery by lexical statistics. *Journal of the American Society for Information Science and Technology*, *50*(1), 574–587.

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

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

Resnick, M., Myers, B., Nakakoji, K., Shneiderman, B., Pausch, R., Selker, T., & Eisenberg, M. (2005). Design principles for tools to support creative thinking. In *Proceedings of the nsf workshop on creativity support tools* (pp. 25–36).

Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing & Management*, *24*(5), 513–523.

Shneiderman, B. (2007). Creativity support tools: Accelerating discovery and innovation. *Communications of the ACM*, *50*(12), 20–32.

Shneiderman, B. (2009). Creativity support tools: A grand challenge for HCI researchers. In M. Redondo, C. Bravo, & M. Ortega (Eds.), *Engineering the user interface: From research to pratice* (pp. 1–9). Springer.

Sluban, B., Juršič, M., Cestnik, B., & Lavrač, N. (2012). Exploring the power of outliers for cross-domain literature mining. In M. Berthold (Ed.), *Bisociative knowledge discovery* (pp. 325–337). Springer.

Smalheiser, N., & 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.

Srinivasan, P. (2004). Text mining: Generating hypotheses from MEDLINE. *Journal of the American Society for Information Science and Technology*, *55*(5), 396–413.

Swanson, D. R. (1988). Migraine and magnesium: Eleven neglected connections. *Perspectives in Biology and Medicine*, *78*(1), 526–557.

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

Swanson, D. R., Smalheiser, N. R., & 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.

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