Expert Systems With Applications 85 (2017) 386–396

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Outlier based literature exploration for cross-domain linking of Alzheimer's disease and gut microbiota

Donatella Gubiani[a,*], Elsa Fabbretti[a], Bojan Cestnik[b,c], Nada Lavrač[b,a], Tanja Urbančič[a,b]

[a] University of Nova Gorica, Nova Gorica, Slovenia
[b] Jožef Stefan Institute, Ljubljana, Slovenia
[c] Temida d.o.o., Ljubljana, Slovenia

ARTICLE INFO

Article history:
Received 1 February 2017
Revised 7 May 2017
Accepted 9 May 2017
Available online 11 May 2017

Keywords:
Literature-based discovery
Outlier detection
Alzheimer's disease
Gut microbiome

ABSTRACT

In knowledge discovery, experts frequently need to combine knowledge from different domains to get new insights and derive new conclusions. Intelligent systems should support the experts in the search for relationships between concepts from different domains, where huge amounts of possible combinations require the systems to be efficient but also sufficiently general, open and interactive to enable the experts to creatively guide the discovery process. The paper proposes a cross-domain literature mining methodology that achieves this functionality by combining the functionality of two complementary text mining tools: clustering and topic ontology creation tool OntoGen and cross-domain bridging terms exploration tool CrossBee. Focusing on outlier documents identified by OntoGen contributes to the efficiency, while CrossBee allows for flexible and user-friendly bridging concepts exploration and identification. The proposed approach, which is domain independent and can support cross-domain knowledge discovery in any field of science, is illustrated on a biomedical case study dealing with Alzheimer's disease, one of the most threatening age-related diseases, deteriorating lives of numerous individuals and challenging the ageing society as a whole. By applying the proposed methodology to Alzheimer's disease and gut microbiota PubMed articles, we have identified Nitric oxide synthase (NOS) as a potentially valuable link between these two domains. The results support the hypothesis of neuroinflammatory nature of Alzheimer's disease, and is indicative for the quest for identifying strategies to control nitric oxide-associated pathways in the periphery and in the brain. By addressing common mediators of inflammation using literature-based discovery, we have succeeded to uncover previously unidentified molecular links between Alzheimer's disease and gut microbiota with a multi-target therapeutic potential.

1. Introduction

Generating new scientific hypotheses has always required a lot of expert knowledge and creativity. Nowadays, it also requires the ability to search, analyse and relate enormous quantities of potentially relevant pieces of information, dispersed in scientific literature and all too often confined to isolated knowledge silos of individual scientific disciplines. Therefore, computational support to scientific discovery, one of the earliest and most successful areas of artificial intelligence (AI) research, has again raised attention of numerous researchers.

In addition to surveying historical examples, including early expert systems such as DENDRAL (Feigenbaum, Buchanan, & Lederberg, 1971), Pat Langley (2000) discusses different steps of the discovery process, from problem formulation to filtering and interpretation, which are needed for a discovery to be accepted in the scientific community. Since then, many approaches to computationally supported knowledge discovery have been developed, including effective literature-based cross-domain knowledge discovery methods (Swanson, 2008). Langley's arguments, including those describing the important role of human developers and users of knowledge discovery methods, are still perfectly valid and are echoed in the essay "The place of literature-based discovery in contemporary scientific practice" by Smalheiser and Torvik (2008). As stated by Swanson (2008), it is important to support and enhance the human knowledge discovery ability with effective computer supported tools that are needed to distinguish potentially interesting new hypotheses from huge amounts of all other possibilities. In this sense, we can see literature-based discovery methods as a potential building blocks of interactive recommender systems (He, Parra, & Verbert, 2016), in which the user's ability to control

* Corresponding author.
E-mail addresses: donatella.gubiani@ung.si (D. Gubiani), e.fabbretti.bf@gmail.com (E. Fabbretti), bojan.cestnik@ijs.si (B. Cestnik), nada.lavrac@ijs.si (N. Lavrač), tanja.urbancic@ung.si (T. Urbančič).

http://dx.doi.org/10.1016/j.eswa.2017.05.026
0957-4174/

the discovery process is one of the important preconditions for computational tools acceptance, sometimes even more important than the accuracy of the used algorithms.

Literature-based discovery (LBD) is a computer supported approach used to identify implicit knowledge from literature databases with the aim of suggesting and supporting new hypotheses, usually by uncovering hidden links between concepts in diverse, previously unconnected scientific literatures. Swanson (1986) was the first to propose text mining approaches to detect cross-domain links via bridging terms (B-terms), connecting previously unrelated medical literature domains. His idea of discovering new hypotheses by connecting fragmented pieces of knowledge from different contexts via bridging terms has proved to be very powerful and has inspired many researchers. Through further development, LBD has matured as a research field on its own, increasingly connected with scientific practice (Bruza & Weeber, 2008) and with a number of successful applications, especially in biomedicine as demonstrated by numerous examples (Erhardt, Schneider, & Blaschke, 2006; Jensen, Saric, & Bork, 2006; Kumar & Tipney, 2014; Oh & Deasy, 2016; Rajpal, Qu, Freudenberg, & Kumar, 2014; Swanson, 1990; Zhang, Sarkar, & Chen, 2014). Several tools have been proposed to exploit LBD methodologies to support experts in the complex task of discovering hidden cross-domain connections, such as ARROWSMITH (Smalheiser & Swanson, 1998), LitLinker (Yetisgen-Yildiz & Pratt, 2006), BITOLA (Hristovski, Peterlin, Mitchell, & Humphrey, 2005), and Literaby (Weeber, 2007).

Given a rapid growth of scientific literature, one of the main problems in LBD is the size of the search space that can be huge. Different approaches cope with this problem in different ways. Many of them use MeSH (Medical Subject Headings) concepts and Unified Medical Language System (UMLS) semantic types to filter the candidates, some examples including (Yetisgen-Yildiz & Pratt, 2006) and (Chen, Lin, & Yang, 2011). Systems that are specialized for a particular type of tasks may use specific background knowledge, for example a thesaurus of gene and protein symbols as used in (Hristovski et al., 2005). On the other hand, if we want the system to be applicable in a wide range of different tasks, more general approaches should be used, one possibility being the use of advanced natural language processing (NLP) approaches. For example, Hristovski, Friedman, Rindflesch, and Peterlin (2008) enhance LBD by capturing semantic relations from the literature with two NLP systems coupled with their LBD system BITOLA.

In this paper we explore NLP in another way. We suggest a new LBD methodology based on the observation that bridging terms indicating potential cross-domain links are more frequent in the outlier documents. In the context of LBD, these are the documents that lie outside the main group of documents of its own domain (Petrič, Cestnik, Lavrač, & Urbančič, 2012; Sluban, Juršič, Cestnik, & Lavrač, 2012). We propose a method that—by combining an outlier detection process with the cross-domain exploration—reduces the set of the documents to be checked by focusing on those documents with a higher probability of containing interesting bridging terms that represent potential cross-domain connections.

We test the proposed methodology exploiting the capabilities of two different software tools—OntoGen for outlier detection and CrossBee for cross-domain exploration—by applying it to the investigation of Alzheimer's disease, one of the most studied neurodegenerative diseases. We concentrate on the "gut-brain axis" with the aim of contributing to a better understanding of Alzheimer's disease, by investigating the links it might have with gut microbiota. Ageing related pathologies, such as Alzheimer's disease and neurodegenerative diseases in general, are a big social and economic problem, leading to numerous societal challenges. Neurodegenerative diseases severely deteriorate lives of many individuals. With ageing of population, they become an urgent priority also due to their social and economic implications. While single-cell mechanisms of ageing processes have been extensively studied, limited knowledge is available on the changes occurring at tissue, organ and system levels leading to the progression of complex chronic age-related disorders, to delineate new hypotheses for potential therapeutic interventions.

Recent clinical literature on gut microbiota supports the strong relationship between the human digestive system and neurodegenerative diseases (such as Alzheimer's disease) indicating new trajectories for original biomedical research (Ghaisas, Maher, & Kanthasamy, 2016). A growing number of scientific articles in this field, including our own research, indicate that a link between dietary and gastrointestinal system and Alzheimer's disease is worth investigating (Gubiani, Petrič, Fabbretti, & Urbančič, 2015), providing a motivation for using text and literature mining methods to identify new hypotheses that can be associated with memory, cognitive dysfunction and brain diseases. This research explores the power of LBD as means for connecting gut microbiota with neurodegenerative diseases, focussing on the "gut-brain axis" with the aim of discovering new potential links between neuronal diseases and gut microbiome. Consequently, in our study we concentrate on the "gut-brain axis" with the aim of contributing to a better understanding of Alzheimer's disease by investigating the links it might have with gut microbiota.

The paper is organized as follows. After explaining the background and motivation for this research in Section 2, Section 3 outlines the proposed methodology for effective cross-domain literature exploration. Section 4 presents the results: the uncovered candidate bridging terms connecting the Alzheimer's disease literature and the gut microbiota literature. Finally, in Section 5 we provide additional connections with the related literature and comment on the application of the methodology by summarizing the results.

2. Background and motivation

This research is motivated by the early work of Swanson (1990) and Smalheiser and Swanson (1998), who developed an approach to assist the user in LBD by detecting interesting cross-domain terms with a goal to uncover new relations between previously unrelated concepts. Their approach has been implemented in online system ARROWSMITH, developed by Smalheiser and Swanson (1998). ARROWSMITH takes as input two sets 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* are further investigated by the user for their potential to generate new scientific hypotheses. 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, 2001), where two initially separate domains *A* and *C* are specified by the user at the beginning of the discovery process, and the goal is to search for bridging concepts (terms) *b* in order to support the validation of the hypothesized connection between *A* and *C*.

The methodology presented in this work upgrades our previous LBD approaches. In the work of Juršič, Cestnik, Urbančič, and Lavrač (2012a), we followed the basic idea of Swanson's "ABC model of knowledge discovery", also aiming at discovering bridging terms *b* as potential links pointing towards new scientific hypotheses. As the identification of bridging terms with high potential relevance for interesting discoveries is a complex process, we based our exploration on using new heuristics that are capable of detecting and ranking the potential bridging terms, where a system of ranking candidate bridging terms by ensemble voting of heuristics was proposed and validated. This methodology was implemented in a user-friendly web application named CrossBee (Cross-Context Bisociation Explorer, (Juršič, Cestnik, Urbančič, & Lavrač, 2012b)).

Table 1
Examples of arguments connecting migraine literature and magnesium literature via bridging concepts (in bold), summarized from Swanson (1990).

Argument 1 Literature on migraine	Argument 2 Literature on Magnesium
Ion channels are involved in migraine attacks.	Magnesium is a **ion channel blocker**.
Stress and Type A behaviour are associated with migraine.	**Stress and Type A behaviour** lead to body loss of magnesium.
Migraine may involve sterile **inflammation**.	Magnesium modulates **inflammation**.
. . .	. . .

The user starts literature-based discovery by uploading documents from two distinct domains, followed by document exploration using the discovered bridging terms (B-terms), ranked by the ensemble of heuristics. Based on our previous findings that outlier documents contain a substantially larger amount of bridging terms than regular (non-outlier) documents (Sluban et al., 2012), some of the CrossBee heuristics have been designed to effectively discover B-terms in the outlier documents. Different options can be set to configure the CrossBee discovery process, where supplementary functionalities and various visualizations help the user to effectively perform cross-domain document exploration.

Following the work of Petrič et al. (2012) and Sluban et al. (2012), this paper approaches cross-domain knowledge discovery by first determining outlier documents for the two scientific domains of interest, followed by the search for particular B-terms. The underlying reasoning is as follows: while the majority of articles in a given specialized scientific domain describe the phenomena related to a common understanding and most intensively investigated issues in the given domain of interest, the exploration of outlier documents may lead to the detection of scientifically, pharmacologically or clinically relevant bridging concepts among sets of scientific articles from two disjoint domains in a novel, not yet explored way. By applying a new outlier-based methodology presented in this paper, we have succeeded to further reduce the set of outlier documents, thus increasing the efficiency and the effectiveness of the knowledge discovery process, given the reduced size of the corpora under investigation that became manageable for expert's inspection and hypothesis formation.

2.1. Medical motivation and research aims

This research is motivated by the early work of Swanson (1990) and Smalheiser and Swanson (1998), who developed an approach to assist the user in LBD by detecting interesting cross-domain terms with a goal to uncover the possible relations between previously unrelated concepts. Swanson's seminal work from more than 25 years ago 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 discovered several connections supportive for the hypothesis that magnesium deficiency might cause migraine headache (Swanson, 1988; 1990; Swanson, Smalheiser, & Torvik, 2006).

Table 1 presents some of the Swanson's examples of discovered bridging concepts and the respective arguments from documents of the two domains that indicate that the discovered terms may indeed be considered as potential bridging terms providing meaningful links between the two domains. Swanson's literature mining results have been 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 other studies, including our previous work (Juršič et al., 2012a).

Recent advancements in understandings human metabolic pathways have highlighted the importance of the function of gut microbiota diversity for human health. Recent data indicate that it represents a key element involved in transformation and absorption of nutrients, immunology balance and integrity of the "gut-brain axis" via proper functioning of the immune system and autonomous nervous system. Increasing knowledge in ageing neuronal pathologies is also recently focusing on impact of active food, as well as in nutrition or malnutrition problems incurring in several categories of patients. The link between "ageing" and "food" knowledge domains was recently studied in information technology (IT) terms (Gubiani et al., 2015) where—starting from literature about gut "microbiota" (microbes that colonize the human gut)—we found the connections of this literature with the literature on Alzheimer's disease, associated to abnormal brain function, and markers of neuronal disorders such as ("homocysteine"), mechanisms involved in protein quality control ("ubiquitin") and markers of synaptic function and learning ("BDNF").

Motivated by this line of research, this paper addresses the LBD problem focusing on potential new discoveries in "gut-brain axis" exploration. Following the ABC model of knowledge discovery, we explore the setting where A corresponds to recent literature on gut microbiota and C corresponds to the literature on Alzheimer's disease.

2.2. IT motivation and research aims

In statistics, an outlier is defined as an observation that falls outside the overall pattern of a distribution (Moore, McCabe, & Craig, 2007). Usually, the presence of outliers is due to data measurement errors and they are discarded. In the area of literature mining, outlier documents are used in a nonstandard text mining task of cross-context link discovery. Sluban et al. (2012) showed that the majority of bridging terms can be found in outlier documents and proved experimental evidence with tests in the gold standard migraine-magnesium domain pair, for which a confirmed list of concept bridging terms was available.

Classification algorithms are one of the techniques that can be used to detect outlier documents (Sluban et al., 2012). Documents from two domains of interest can be used to train a classification model that distinguishes between the documents of these two domains. The constructed model allows one to classify all the documents and, those that are misclassified are declared as outlier documents, since according to the classification model they do not belong to their original domain. The model considers them to be more similar to the documents of the other domain than to the documents of their originating domain. 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 we denote a set of such outlier documents with $\mathcal{O}(A)$. Similarly, the set of documents originally from the class C but classified by a classification algorithm into class A is denoted as $\mathcal{O}(C)$.

Sluban et al. (2012) tested the hypothesis that domain outliers obtained by classification noise detection have the potential for bridging different concepts. This hypothesis was tested on

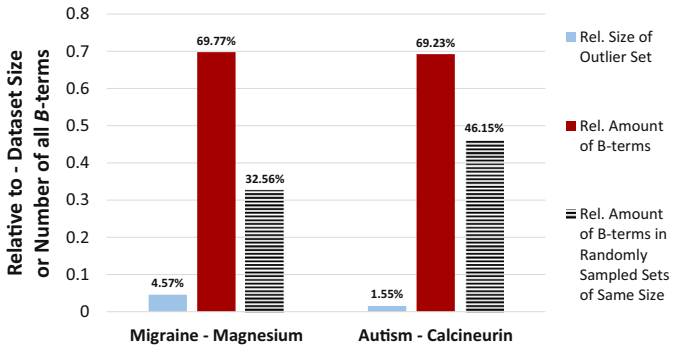


Fig. 1. *B*-terms in the detected outlier sets of two domain pair datasets. The experimental results on the migraine–magnesium (Swanson et al., 2006) and on the autism–calcineurin (Petrič et al., 2009) datasets show that the sets of detected outlier documents were relatively small (less than 5% of the entire datasets) and that they contained a great majority of bridging terms (around 69%), which was significantly higher than in same-sized random subsets (around 32% and 46%).

the migraine–magnesium (Swanson et al., 2006) and the autism–calcineurin (Petrič, Urbančič, Cestnik, & Macedoni-Lukšič, 2009) domain pair datasets, with lists of confirmed concept bridging terms (B-terms) available for testing. 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 bridging terms, which was significantly higher than in same-sized random subsets. These results, summarized in Fig. 1, indicate that the effort needed for finding cross-domain links can be substantially reduced due to exploring a much smaller subset of outlier documents, where a great majority of *B*-terms are present and more frequent.

A different approach to outlier document detection is by using clustering algorithms. Following Petrič et al. (2012), this work uses the OntoGen document clustering tool (Fortuna, Grobelnik, & Mladenić, 2006) to find outliers, focusing on domain outlier documents that tend to be more similar to the documents of the opposite domain than to those of their own domain. OntoGen supports unsupervised or supervised document clustering. The unsupervised algorithm is based on *k*-means clustering (Jain, Murty, & Flynn, 1999). First, the OntoGen's 2-means clustering algorithm is applied to cluster the merged document set $A \cup C$ (labelled `root` in Fig. 2). The result of this unsupervised clustering is a set of two document clusters: A' (labelled `Classified as` A in Fig. 2) consists mainly of documents from A, but may contain also some documents from C, and similarly C' (labelled `Classified as` C in Fig. 2) consists mainly of documents from C, but may contain also some documents from A. Then, for each of the clusters, a supervised clustering approach is applied taking into account the documents' original domains A and C. As a result, a two-level tree hierarchy of clusters is generated (Fig. 2) and, at the second level, we can identify outliers $\mathcal{O}(A)$ and $\mathcal{O}(C)$ as the documents categorized by 2-means clustering into the other domain than the domain of their origin. A specifically interesting feature of OntoGen for outlier document detection is similarity graph visualization for two given document sets (e.g., $A \cup C$), constructed by ranking and visualizing all the documents in terms of their similarity to the centroid of each individual document set (e.g., similarity to centroid c_A of A and similarity to centroid c_C of C, respectively). A similarity graph is illustrated in Fig. 8 in Section 4.

Motivated by the described line of research, this paper proposes an extended methodology based on the detection and exploration of outlier documents, aimed at uncovering new, previously unidentified molecular "gut-brain axis" links.

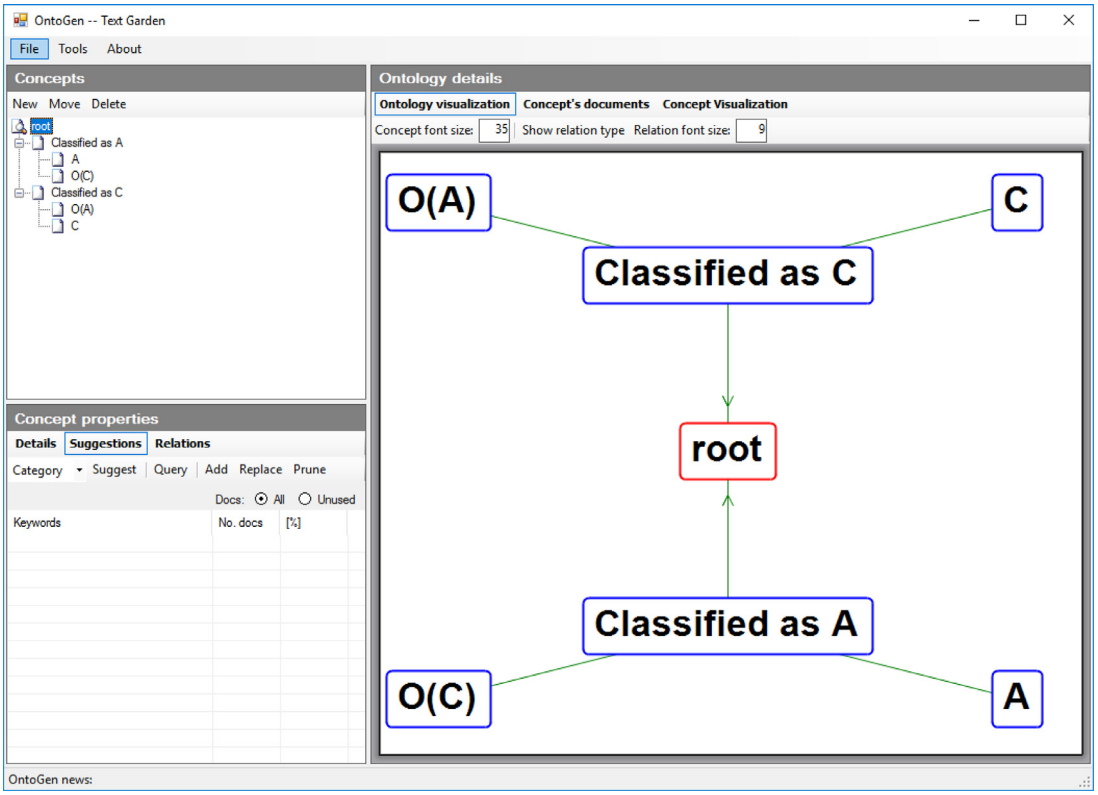


Fig. 2. Outlier detection in OntoGen. Target domain documents from literatures *A* and *C* are clustered according to the OntoGen's two step approach to obtain outlier documents $\mathcal{O}(A)$ and $\mathcal{O}(C)$: first using unsupervised and then supervised clustering.

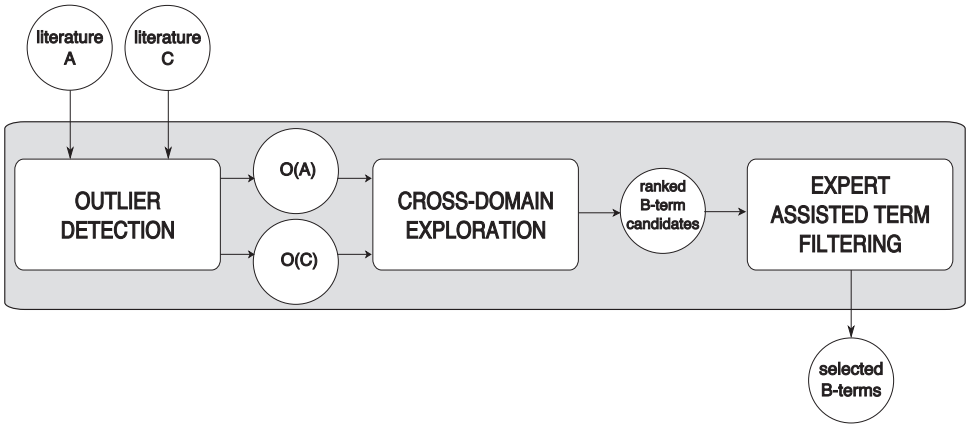


Fig. 3. Schematic overview of the proposed methodology. Starting from two selected sets of documents (literatures A and C), an outlier detection step is used first to select sets of outlier documents $\mathcal{O}(A)$ and $\mathcal{O}(C)$. These sets become an input for cross-domain exploration step in which candidates for bridging terms are identified and ranked. Finally, with the expert assistance, the list of candidates is further filtered to select bridging terms supporting a new scientific hypothesis to be checked with methods commonly accepted in the biomedical research community.

3. Methodology

In order to increase the efficiency and effectiveness of the knowledge discovery process, we developed a new LBD methodology for detecting hidden connections between distinct literature domains, by reducing the investigation only to outlier documents instead of considering entire literatures. For this reason, we upgraded the CrossBee methodology developed by Juršič et al. (2012a) with the exploration of outlier literatures as proposed by Petrič et al. (2012). The proposed methodology is schematically depicted in Fig. 3.

3.1. Overview of the proposed methodology

The proposed methodology works on two input literatures (*literature A* and *literature C*) that can be retrieved from a bibliographic database, such as PubMed (https://www.ncbi.nlm.nih.gov/pubmed/). The methodology consists of the three steps described below.

Step 1: Outlier detection. We adopt the methodology proposed by Petrič et al. (2012) to a two-step outlier detection process supported by the OntoGen document clustering tool (Fortuna et al., 2006). The documents from the two individual sets are loaded as a single text file (i.e. a joint document set named $AC = A \cup C$) in which each document is identified by the *PMID* (PubMed IDentifier) and described by domain label (A or C), the title and the abstract (we used titles and abstracts based on previous experimental evidence by Petrič, Urbančič, & Cestnik (2006)).

An upgraded two-level clustering approach is used to detect outlier documents. At the first level, the result of unsupervised clustering based on two document clusters determines A' (i.e. the set of documents from $A \cup C$ classified as A), and C' (i.e. the set of documents from $A \cup C$ classified as C). Then, at the second level, each of these clusters is further divided into two document subclusters based on domain labels (A or C) with the aim to identify outlying documents: cluster A' is divided into sub-clusters $A' \cap A$ and $A' \cap C$, while cluster C' is divided into $C' \cap A$ and $C' \cap C$. In this manner, sub-cluster $A' \cap C$ determines outliers of C (denoted as $\mathcal{O}(C)$), consisting of those documents that were obtained originally as members of C but are now classified into A' since they were recognized to be more similar to the documents from A than to the ones from C. Similarly, $C' \cap A$ is a set of outliers of A (denoted as $\mathcal{O}(A)$), consisting of the documents obtained originally from the domain A but classified into C' based on their similarity with the documents from domain C.

Step 2: Cross-domain exploration. In this step, bridging terms are searched for by CrossBee, a user-friendly web application that implements the methodology for cross-domain exploration developed by Juršič et al. (2012a). The outlier documents $\mathcal{O}(A)$ and $\mathcal{O}(C)$ detected by OntoGen are loaded in CrossBee as a single joint text file $\mathcal{O}(A) \cup \mathcal{O}(C) = \mathcal{O}(A)\mathcal{O}(C)$. This input file is similar to the one for OntoGen: it only differs for separator symbols and the source of data (PubMed export file vs. OntoGen export file). CrossBee suggests a list of B-terms by using an ensemble heuristics, combining different functions measuring the likelihood of a term being a B-term. The output is a ranked list of B-term candidates, the ones with the highest score as a result of the ensemble heuristics being at the top of the list. Heuristics are advanced term statistics (Juršič et al., 2012a), which are either frequency based, TF-IDF weights-based, similarity based and outlier based evaluations of the given term in a particular document set.

Step 3: Expert assisted term filtering. In an ideal scenario with a perfect ensemble heuristics, all B-terms should be at the top of the ranked CrossBee list of candidate bridging terms. Although we want to come as close to this goal as possible, this is not realistic in the current framework and additional filtering of candidates is needed. This is done in two steps: (3.1) discarding the terms that have already been studied in both investigated literatures, and (3.2) further term filtering based on expert's suggestions, allowing the system to focus on new potentially interesting links aligned with expert's specific research interests. While step 3.1 is fully automated (with using combined queries to check combinations in all articles available in PubMed), step 3.2 is done in a close interaction with the expert. This way, out of different links between the investigated domains, those with the highest matching with the expert's intuition and interests will be addressed. Experts can decide for MeSH filtering (implemented also in CrossBee), but—as in our case—they might find this option too restrictive and prefer their own view on the domain, focusing on the particular features they are interested in. In the beginning of this process, some of the terms can be filtered out easily by not being in a category (could also be a particular MeSH category) declared as relevant by the expert. In the continuation, having a list of candidate terms together with CrossBee's additional functionalities for visualizations of terms and documents helps the expert to investigate and narrow down the list of candidates. In particular, it is beneficial to have a

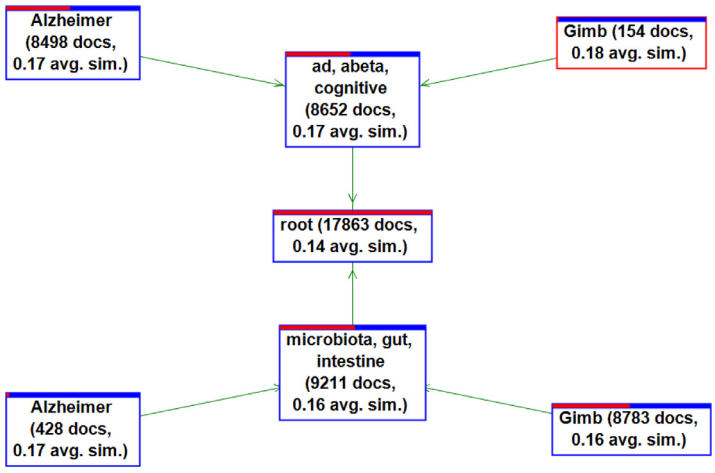


Fig. 4. Two-level cluster hierarchy with OntoGen. The ontology constructed from 17,863 papers includes two first-level clusters labelled as expected with terms appropriate for the original domains, such as "AD, Aβ, cognitive" and "microbiota, gut, intestine" for Alzheimer and GIMB literatures, respectively. Four second level sub-clusters separate documents according to their original search keyword.

set of recommended documents to be checked in each particular case.

As a remark, note that the proposed methodology could be used also if domain C were not pre-specified by the user. In this case, one could use an open discovery process to identify domain C, using the RaJoLink method and the corresponding software tool, developed by Petrič et al. (2009). RaJoLink allows for starting from a specific literature A (associated to domain A), to determine a candidate domain C by exploring rare terms r, and the intersection of their corresponding literatures B.

In the next two sections, we describe the materials and present the application of the methodology in a case study of linking the Alzheimer's disease and gut microbiota literature. In particular, we give more details about the three methodological steps as used in this application.

3.2. Datasets

In our study, the set of Alzheimer's disease documents consisted of 83,322 documents and was obtained from PubMed by posing the query "Alzheimer". For gut microbiota, in order to work with similar numbers of elements in the two domains, we used a more elaborate query "(gut OR intestinal) AND (microbiota OR bacteria)" (referred to as GIMB in this paper). The resulting set for GIMB included 73,960 documents. However, due to the functional limitations of the tool, these sets were reduced by eliminating documents with incomplete title or abstract, and by restricting the scope of documents to those published in years 2014 and 2015. The resulting sets are composed of 8,934 documents for "Alzheimer" (domain A) and 8,937 for "GIMB" (domain C).

3.3. Methodology applied to Alzheimer's disease and GIMB literatures

The three steps of the proposed methodology are described below.

3.3.1. Outlier document detection with OntoGen

On the joint set of 17,863 documents ($Alzheimer \cup GIMB$), we generated the two-level document hierarchy with OntoGen (Fortuna, Grobelnik, & Mladenić, 2005), with the aim of getting an insight into the contents structure of the documents and of identifying the outlier documents.

At the first level, after transforming the documents into a feature vector format, the documents were clustered according to their similarity (the "cosine similarity measure" implemented in OntoGen was used) into two distinct document clusters. By checking the most relevant concepts for each cluster, the clusters are clearly associated to the two original two domains. As illustrated in Fig. 4, the first cluster with 8,652 documents is identified by concepts related to Alzheimer's disease, and the second one with 9,211 documents is related to the GIMB domain.

At the second level, each of the two clusters was further separated according to the documents' search origin ("Alzheimer" or "GIMB") into two sub-clusters. The constructed hierarchy in Fig. 4 shows how we got 582 outlier documents: 428 from "Alzheimer" ($\mathcal{O}(Alzheimer)$ documents were automatically classified as "GIMB" although they originated from the "Alzheimer" domain) and 154 from "gut microbiota" ($\mathcal{O}(GIMB)$ documents were automatically classified as belonging to domain "Alzheimer", although in origin they were from the "GIMB" domain).

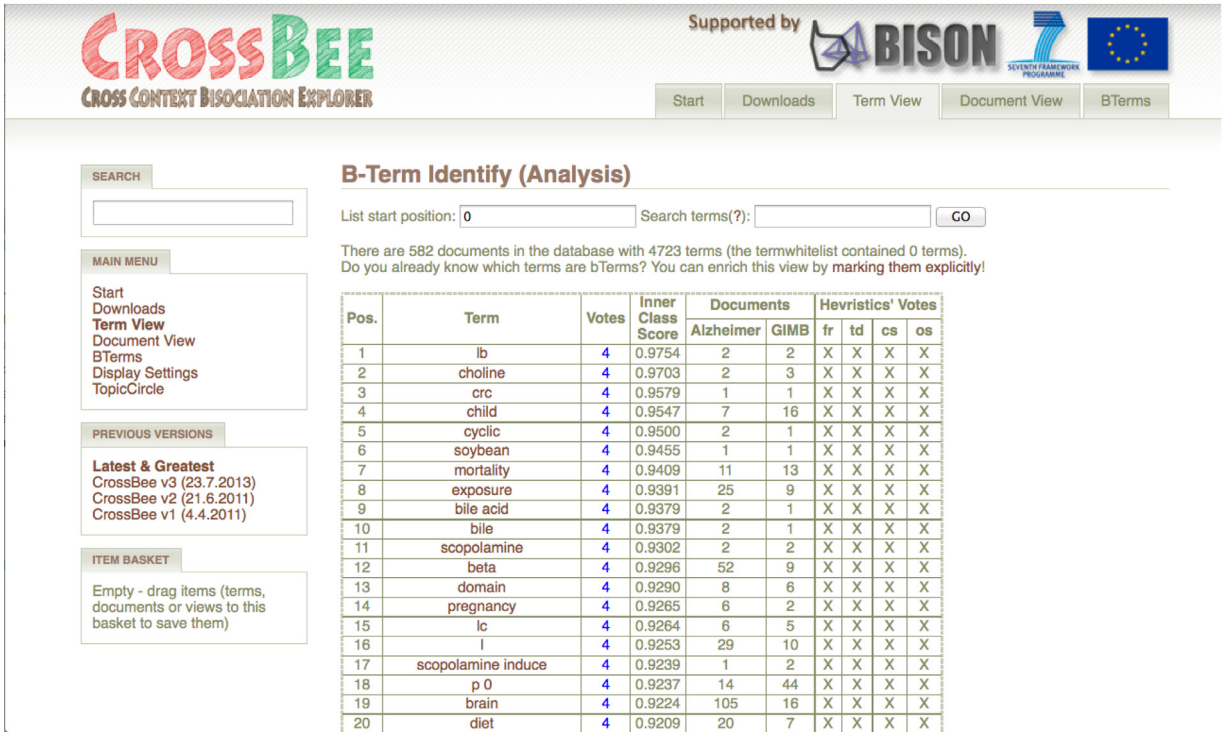
3.3.2. Cross-domain exploration with CrossBee

We further explored the 582 outlier documents using the CrossBee tool (Juršič et al., 2012b). By processing the document set $\mathcal{O}(Alzheimer)\mathcal{O}(GIMB)$, CrossBee extracted a list of 4,723 terms as potential bisociative terms linking the two analysed domains. The top of the ranked list of potential B-terms is shown in Fig. 5. The terms are ranked according to the estimated bridging term potential as proposed in (Juršič et al., 2012b) and each term is associated with the frequency in the outlier document sets.

3.3.3. Expert assisted term filtering

First, automatic filtering was applied to the list of 4,723 terms identified by CrossBee. This filter excluded the terms that were already studied in both literatures (in this step, checking all the available documents and not just the outliers, and not restricted by the year of publication). This has reduced the list to 2,513 terms. Next, the expert excluded terms that do not belong to the domain-specific biomedical terminology, such as numbers, measures, verbs, etc. With this step, the set of terms was reduced to 572.

In the continuation, the expert analysed these terms by grouping them into 5 categories: (i) Chemicals, mechanisms of action, cell components (201 terms); (ii) Diseases, organs, tissues (204 terms); (iii) Biological agents, including bacteria and viruses (63 terms); (iv) Genetic mechanisms (27 terms); and (v) Other (76 terms). The expert decided to focus on the first category, as it was the most relevant for the identification of mechanisms of possible pharmacological interest. When this category was further clustered



Pos.	Term	Votes	Inner Class Score	Documents		Hevristics' Votes			
				Alzheimer	GIMB	fr	td	cs	os
1	lb	4	0.9754	2	2	X	X	X	X
2	choline	4	0.9703	2	3	X	X	X	X
3	crc	4	0.9579	1	1	X	X	X	X
4	child	4	0.9547	7	16	X	X	X	X
5	cyclic	4	0.9500	2	1	X	X	X	X
6	soybean	4	0.9455	1	1	X	X	X	X
7	mortality	4	0.9409	11	13	X	X	X	X
8	exposure	4	0.9391	25	9	X	X	X	X
9	bile acid	4	0.9379	2	1	X	X	X	X
10	bile	4	0.9379	2	1	X	X	X	X
11	scopolamine	4	0.9302	2	2	X	X	X	X
12	beta	4	0.9296	52	9	X	X	X	X
13	domain	4	0.9290	8	6	X	X	X	X
14	pregnancy	4	0.9265	6	2	X	X	X	X
15	lc	4	0.9264	6	5	X	X	X	X
16	l	4	0.9253	29	10	X	X	X	X
17	scopolamine induce	4	0.9239	1	2	X	X	X	X
18	p 0	4	0.9237	14	44	X	X	X	X
19	brain	4	0.9224	105	16	X	X	X	X
20	diet	4	0.9209	20	7	X	X	X	X

Fig. 5. Cross-domain exploration with CrossBee. CrossBee identified 4,723 B-terms and displayed them sorted by the ensemble heuristic value (*Inner Class Score*), together with their frequency in the starting outlier sets (*Documents*).

at the next level, a very evident subcluster of terms was related to oxidative stress (31 terms). Among them, "Nitric oxide synthase" was identified as a promising novel bridging term of importance for the neuronal and for the immunity field.

Nitric oxide synthase (NOS) (Katusic & Austin, 2014) is an enzyme responsible for the synthesis of nitric oxide (NO), with a strong role in physiological and pathological conditions, with a modulatory and inflammatory potential. While in normal conditions NOS is expressed in neurons (nNOS) or endothelial cells (eNOS), where it is involved in neuro- and vaso-active effects, its expression is strongly induced during inflammation (iNOS). The inducible form of NOS (iNOS) was a potential bridging term that was chosen because of its unique property to cover multiple fields of interest, namely immunity and inflammation, oxidative stress and neurodegenerative aspects. This finding is coherent with the knowledge that iNOS is an important well-known mediator of brain and gut inflammation pathologies. In particular, microscopy analysis gut biopsies from human patients affected by acute gut inflammation, identified higher iNOS expression respect to healthy control subjects (Kolios, Rooney, Murphy, Robertson, & Westwick, 1998; Middleton, Shorthouse, & Hunter, 1993). iNOS was not found in gut non-inflamed samples. In addition, inflammatory cytokines like IL-1, tumour necrosis factor alpha (TNF-α) and interferon gamma (INF-γ) are involved in induction of iNOS expression. Analyses of gut tissue from patients with inflammatory bowel disease (IBD) have also shown a significant increase in iNOS expression and local NO signalling (Boughton-Smith et al., 1993; Lundberg, Lundbergand, Alving, & Hellstrom, 1994).

To validate the methodology, CrossBee was applied to analyse the bridging term "Nitric oxide synthase", as illustrated in Fig. 6). Using CrossBee, NOS was identified in three documents (Gan et al., 2015; Mancuso & Santangelo, 2014; Rannikko, Weber, & Kahle, 2015) from domain "Alzheimer" and in one document (Xiao et al., 2014) from domain "GIMB". This outcome is depicted in Fig. 7.

4. Results

We have analysed the four documents detected by CrossBee to identity the bridging term "Nitric oxide synthase" (Fig. 6): three papers from Alzheimer literature (Gan et al., 2015; Mancuso & Santangelo, 2014; Rannikko et al., 2015) and one from gut microbiota literature (Xiao et al., 2014). These outlier documents are interesting. As an example, consider the abstract of Paper (Xiao et al., 2014), which is an outlier document in the PubMed gut microbiota literature, as shown in the similarity graph in Fig. 8 as a blue dot outlier among the red Alzheimer's disease document line. From our perspective it is noticeable that the abstract of this article presents a hypothesis that "Learning and memory abilities are associated with alterations in gut function" (Xiao et al., 2014). To verify this hypothesis, the authors used behavioural and neural network experiments to demonstrate a synergistic activity of Lactobacilli and plant anthocyanidins in enhancing learning and memory in animal models. In addition, the expression of NOS as mediator of these processes in brain, serum and colon, was also identified (Xiao et al., 2014).

In our work, by selecting NOS as a bridging term of interest and by further investigating the outlier documents, we found that NOS may indeed act as a yet unexplored connection between the Alzheimer's disease literature and the literature on gut microbiota, even supported by known importance of NO as neurotransmitter in the peripheral and central nervous system as well as its role in inflammation. See the indicative sentences in Table 2, extracted from the outlier documents, which provide arguments for the new medical hypothesis that NOS is indeed an interesting B-term worth investigating as a link between gut microbiota and Alzheimer's disease.

The domain expert has interpreted these findings as follows.

- NOSs (Nitric oxide synthases) generate NO (nitric oxide), which acts as neuronal and inflammatory mediator in the gut and the brain.

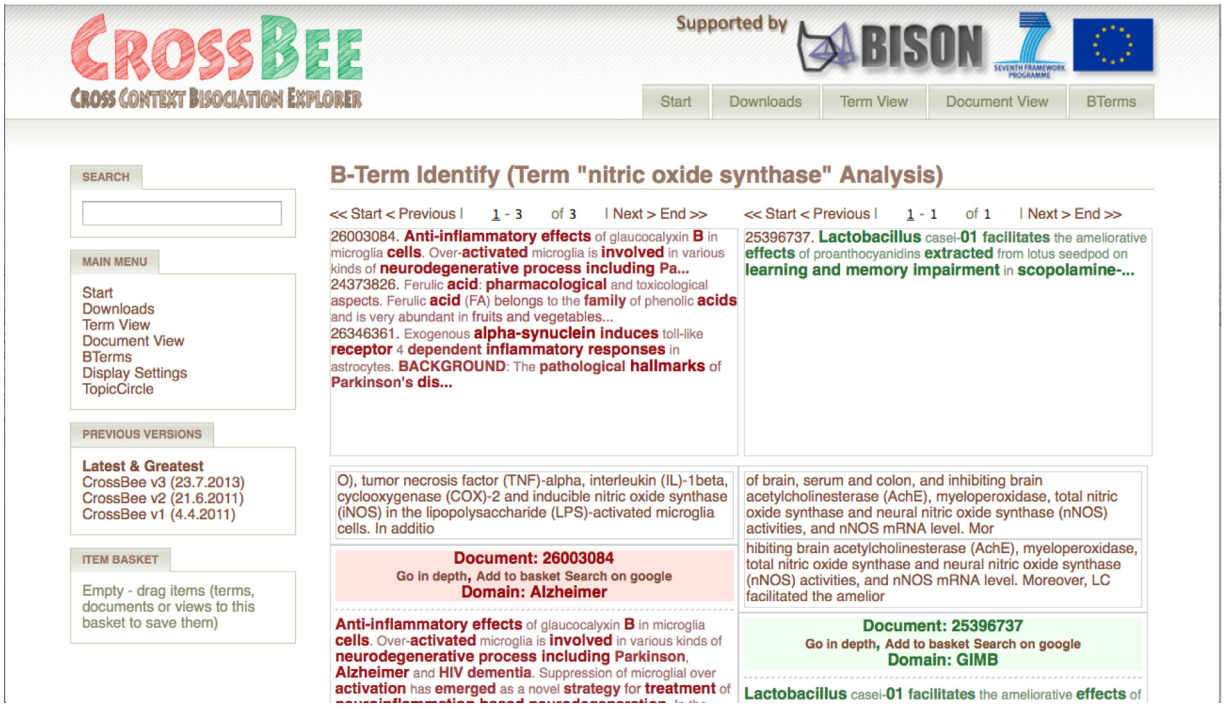


Fig. 6. Analysis of term "nitric oxide synthase" in CrossBee. The bridging term is identified by three papers from Alzheimer literature (Gan et al., 2015; Mancuso & Santangelo, 2014; Rannikko et al., 2015) and one from gut microbiota literature (Xiao et al., 2014).

Table 2
Bridging terms (in bold), discovered as links between the Alzheimer's disease and gut microbiota PubMed literatures.

Argument 1 Literature on Alzheimer's disease	Argument 2 Literature on gut microbiota
NOSs generate NO, which acts as neurotransmitter in the brain. Excess of **NOS** activity is associated to brain diseases and neuroinflammation. **NOS**s in inflammatory cells influence the immune system function. . . .	**NOS**s generate NO, which acts as neurotransmitter in the gut. **NOS** is associated to aberrant gut inflammation. Microbiota dysbiosis induces expression of **NOS** in immune cells. . . .

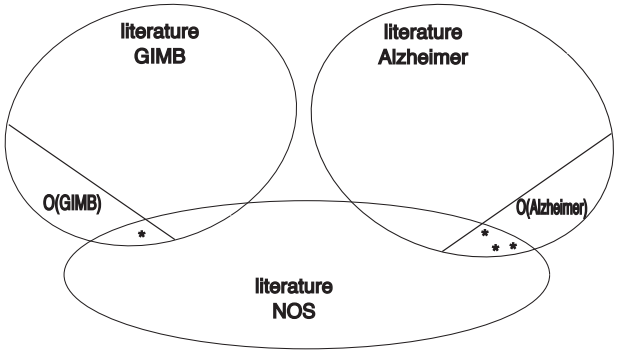


Fig. 7. "Nitric oxide synthase" literature. The figure illustrates that the literature mentioning "Nitric oxide synthase" has an intersection with two individual literatures on Alzheimer's disease and GIMB, while the identified bridging term "Nitric oxide synthase" has not been found in any paper mentioning both Alzheimer's disease and GIMB. The stars represent the four documents identified by CrossBee. As in Fig. 6, these are papers (Gan et al., 2015), (Mancuso & Santangelo, 2014) and (Rannikko et al., 2015) from Alzheimer's and (Xiao et al., 2014) from GIMB literature.

- NO levels are associated to health or disease states; therefore NOS activity is an important parameter.
- Microbiota dysbiosis might imply changes in immune system function, as well as different availability of gut-derived neuroactive molecules (such as serotonin) that strongly influence brain function.

Further work needs to make the link clearer, however, the knowledge today suggests that similar neurodegenerative mechanisms occur in the brain and in the gut of ill elderly, making it worth to identify new druggable targets for both districts of the human body. Argumentation for further research on NOS is provided in Section 5.

5. Discussion and conclusions

In the context of rapid growth of scientific literature, IT based discovery methods can provide useful support to experts in a complex knowledge discovery task of identifying cross-domain links, which may lead to new scientific insights. This paper addresses the problem of effectively reducing a huge search space of possible cross-domain links by combining two different approaches to cross-domain knowledge discovery, showcased in a difficult and challenging problem of finding potentially insightful links explaining joint "gut-brain axis" phenomena.

As the input to the addressed literature-based discovery task, we took published PubMed articles in two distinct literatures: i.e. papers on gut microbiota and Alzheimer's disease. Our research suggests a new hypothesis about the role of NOS/NO in human pathology, discovered using a new combined methodology to find bisociative links through outlier documents. The methodology exploits an interplay of two existing software tools: OntoGen (Fortuna et al., 2005) and CrossBee (Juršič et al., 2012a). By using the OntoGen clustering tool to detect outlier documents, and by using the expert-provided list of terms of potential interest, we

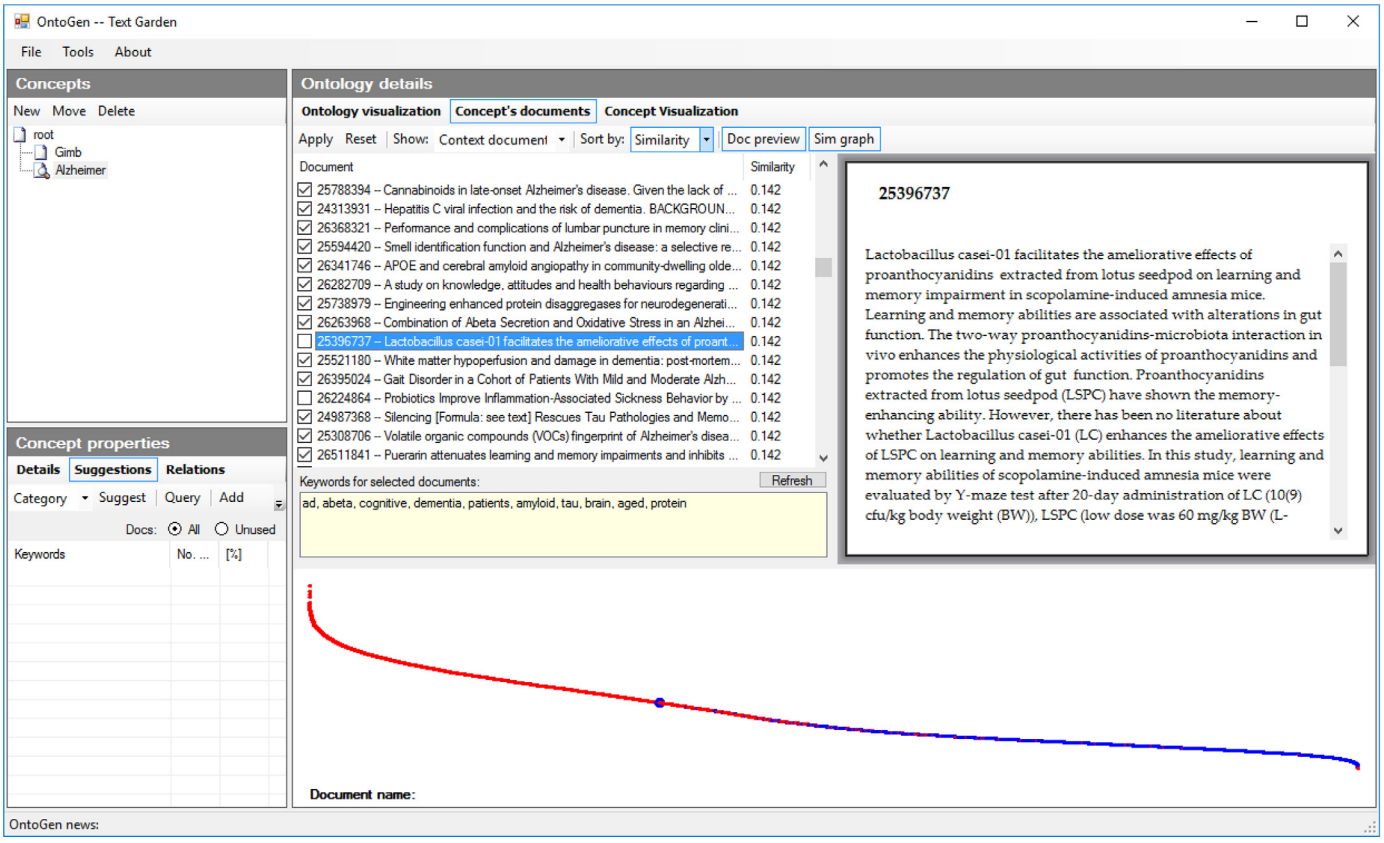


Fig. 8. Outlier document (Xiao et al., 2014) in the similarity graph. Document (Xiao et al., 2014) (PubMed id: 25396737) from the gut microbiota domain is shown in the similarity graph as a blue dot outlier among the red Alzheimer's disease document line.

have succeeded to effectively narrow down the CrossBee search for bridging terms.

Our methodology has proved to be successful in discovering novel and relevant cross-domain links. It is especially useful in cases where the size of investigated domains is a limiting factor, making ranking of potential links too difficult given a huge search space of possible cross-domain connections. One of the main strengths of our approach is the reduction of the search space performed in a general (not domain specific) way by restricting the search space to outlier documents identified with OntoGen. For this reason, our outlier based-approach is effective and could be used practically unchanged also in other, possibly very different problem domains, such as e.g., ecology, where important discoveries arise from the investigation of rare events or conditions of many different types (Ellison & Agrawal, 2005).

Additional advantage is the functionality provided by CrossBee, which offers ranking of candidate bridging terms using ensemble heuristics, and a user-friendly CrossBee's interface with different visualizations of documents and terms supporting experts in making decisions about further narrowing or changing the focus in the next steps of knowledge discovery. This is in line with the Swanson's discussion in (Swanson, 2008), emphasizing the need of computational knowledge discovery tools to support experts by suggesting different variants together with the information about their potential for hypothesis generation, while leaving enough openness for the expert to guide the process according to his or her research interests and intuition. This way, our methodology not just supports the knowledge discovery process by focusing the expert's attention towards the more promising terms, but also by enhancing the expert's genuine creativity. In our practice we have several times witnessed the moments in which the experts very creatively generated their own new ideas triggered by the results of the software tool.

The main weakness—requiring further work—is related to the final choice of candidate bridging terms to be selected for hypothesis testing, since there is still potential to better support the experts in this phase. Although in our approach the idea is not to fully automate the knowledge discovery process but rather to support the experts with a powerful and effective tool providing recommendations for hypotheses generation, the main issue in our future work will be to further decrease the expert's effort in the parts of the process, while still leaving them the opportunity to guide the process in accordance with their research interests.

Existing predefined categorizations, such as MeSH, may be useful for discarding some obviously irrelevant subsets of terms, but it turns out that this is not sufficient since it may not reflect the features perceived as most relevant by the expert when searching for new hypotheses. Also, such predefined general categorizations do not reflect statistical characteristics of a specific set of input documents. To improve this part of the process, we intend to use semi-automated generation of ontologies for each of the two investigated domains and to investigate existing/non-existing links with combined queries for pairs of cluster keywords (one from each of the two investigated domains), identifying "white spots" on a higher abstract level. Prioritizing within a chosen subclass of terms is done based on ensemble heuristics in CrossBee.

Further enhancement of the heuristics included in the ensemble may improve this part of the process. Based on our previous experience we will do this by focusing primarily on outlier based statistics to detect outlier documents, and by using predefined controlled vocabularies as already suggested in (Perovšek, Juršič, Cestnik, & Lavrač, 2016a). Additionally, we believe that terms that are

rare in the investigated context have a potential not yet fully exploited. They have been exploited in the "Ra" step of the RaJoLink approach presented in Petrič et al. (2009) for suggesting the domain to be connected with the investigated domain, while the potential use of rare terms (that could in this sense be viewed as term outliers) in the "Link" step remains to be investigated. A related idea is to use term extraction—e.g., by OKAPI25—to explore and utilize the specificity of a particular term for a particular domain represented with a set of documents.

Note that most of suggested future improvements preserve the general and domain-independent character. We will also put additional attention into the development of the human-computer interface, since we believe that an interactive approach exploiting domain experts' knowledge will remain an important reference for discovery speedup and for the validation of hypothesized discoveries. One of the directions for supporting this interactivity is through implementations in text mining platform TextFlows (Perovšek, Kranjc, Erjavec, Cestnik, & Lavrač, 2016b), already adapted also for knowledge discovery tasks (Cestnik, Fabbretti, Gubiani, Urbančič, & Lavrač, 2017; Perovšek et al., 2016a).

We proceed with a discussion related to the impact of our findings in the biomedical field. In the presented case study, among several candidate bridging terms we focused on "Nitric oxide synthase" as a promising novel bridging term, likely describing a physiological role of nNOS and NO in both brain and gut regions as well as inducibly expressed in different pathological conditions (iNOS). Nitric oxide synthases (NOSs) are enzymes expressed in neurons (nNOS), endothelial cells (eNOS) and in immune cells (iNOS). NO-mediated innervation is found in the gut peripheral nervous system (Phillips & Powley, 2007; Rivera, Poole, Thacker, & Furness, 2011), and in the brain, where NO controls brain regions susceptible to neurodegeneration (Blomeley, Cains, & Bracci, 2015; Steinert, Chernova, & Forsythe, 2010; Toda & Okamura, 2012). From the two sets of documents used as input, we identified the documents that were carrying the chosen bridging term, namely three documents from the Alzheimer's disease domain (Gan et al., 2015; Mancuso & Santangelo, 2014; Rannikko et al., 2015) and one from the gut microbiota domain (Xiao et al., 2014), while the identified bridging term has not been previously explored in the "gut-brain axis" literature. This is evidenced by the fact that the combined query "Alzheimer gut nitric oxide" in PubMed revealed no elements of connection, which suggest the novelty of the discovered link.

In the light of current knowledge published in the PubMed literature, the finding can be interpreted in view of microbiota contribution to iNOS-mediated inflammation at the gut level (Baruch, Kertser, Porat, & Schwartz, 2015; Derkinderen et al., 2011; Xiao et al., 2014). Although not yet demonstrated, it is likely that such effects further dysregulate the brain-gut communication. In pathological conditions, such as during inflammation or in the presence of environmental stressors or ageing, abnormal NO levels results in oxidative effects and neurodegeneration. In particular, in the gut, iNOS induces intestinal barrier damage (Grishin, Bowling, Bell, Wang, & Ford, 2016), and in the brain causes nitrosylation of proteins and cell death with neurological consequences like dementia, Alzheimer's or Parkinson's disease (Hess, Matsumoto, Kim, Marshall, & Stamler, 2005; Horn et al., 2002). Even though the effect of iNOS is local, we cannot exclude that its role in nitrergic gut and brain neurons can be similar, as suggested, therefore influencing the progression of the disease (Gan et al., 2015; Mancuso & Santangelo, 2014; Rannikko et al., 2015). Furthermore, the potential of personalized medicine, smartfood, and microbiome-based therapeutics are of great interest today. Bioactive nutrients possibly modulate individual microbiota responses limiting inflammation and stress responses, including NO (Jeong et al., 2015). Although NO/iNOS targeted therapeutic strategies were already proposed (Broom et al., 2011), further studies are necessary to clarify the consequences of pathological NO signalling in different tissues.

How gut microbiota influences the brain function is a matter of intense studies and the molecular link between gut and brain within the "gut-brain axis" is not known. Scientific and medical literature however is supportive of effective crosstalk between the two compartments. The final clarification of the role of iNOS in dysbiosis of gut microbiota requires a further validation in experimental models of Alzheimer's disease. Nevertheless, our results demonstrate the utility of a search engine method for generating new research hypotheses to drive new research approaches or to identify new druggable targets.

In summary, our work proved to be effective in identifying common molecular targets that have a role in modulating the microbiota/gut/brain axis, supporting the interest for multi-purpose therapeutic strategies, able to contain oxidative and inflammatory processes with high relevance for peripheral and brain neuron function.

Acknowledgments

We acknowledge the European Commission's support through the Human Brain Project (grant no. 604102), and support of the Slovenian Research Agency program Knowledge Technologies and project Analysis of heterogeneous information networks for knowledge discovery in life-sciences. We are grateful to Borut Sluban and Ingrid Petrič for contributing to the research on outlier detection.

References

Baruch, K., Kertser, A., Porat, Z., & Schwartz, M. (2015). Cerebral nitric oxide represses choroid plexus NFκB-dependent gateway activity for leukocyte trafficking. *The EMBO Journal, 34*(13), 1816–1828.

Blomeley, C. P., Cains, S., & Bracci, E. (2015). Dual nitrergic/cholinergic control of short-term plasticity of corticostriatal inputs to striatal projection neurons. *Frontiers in Cellular Neuroscience, 9*, 453.

Boughton-Smith, N. K., Evans, S. M., Hawkey, C. J., Cole, A. T., Balsitis, M., Whittle, B. J., et al. (1993). Nitric oxide synthase activity in ulcerative colitis and Crohn's disease. *The Lancet, 342*, 338–340.

Broom, L., Marinova-Mutafchieva, L., Sadeghianand, M., Davis, J. B., Medhurst, A. D., & Dexter, D. T. (2011). Neuroprotection by the selective iNOS inhibitor GW274150 in a model of Parkinson disease. *Free Radical Biology and Medicine, 50*(51), 633–640.

Bruza, P., & Weeber, M. (2008). *Literature-based discovery* (1st). Springer Publishing Company, Incorporated.

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

Chen, R., Lin, H., & Yang, Z. (2011). Passage retrieval based hidden knowledge discovery from biomedical literature. *Expert Systems with Applications, 38*, 9958–9964.

Derkinderen, P., Rouaud, T., Lebouvier, T., Bruley des Varannes, S., Neunlist, M., & De Giorgio, R. (2011). Parkinson disease: the enteric nervous system spills its guts. *Neurology, 77*(19), 1761–1767.

Ellison, A. M., & Agrawal, A. A. (2005). The statistics of rarity. *Ecology, 86*, 1079–1080.

Erhardt, R. A.-A., Schneider, R., & Blaschke, C. (2006). Status of text-mining techniques applied to biomedical text. *Drug Discovery Today, 11*(7/8), 315–325.

Feigenbaum, E. A., Buchanan, B. G., & Lederberg, J. (1971). On generality and problem solving: A case study using the DENDRAL program. In B. Meltzer, & D. Michie (Eds.), *Machine intelligence: 6* (pp. 165–190). Edinburgh University Press.

Fortuna, B., Grobelnik, M., & Mladenić, D. (2005). Semi-automatic construction of topic ontologies. In *Proceedings of joint international workshops on semantics, web and mining (EWMF 2005 and KDO 2005)* (pp. 121–131). http://ontogen.ijs.si/. Accessed 07.05.17.

Fortuna, B., Grobelnik, M., & Mladenić, D. (2006). Semi-automatic data-driven ontology construction system. In *Proceedings of the 9th international multi-conference information society (IS 2006)* (pp. 223–226). http://ontogen.ijs.si/. Accessed 07.05.17.

Gan, P., Zhang, L., Chen, Y., Zhang, Y., Zhang, F., Zhou, X., et al. (2015). Anti-inflammatory effects of glaucocalyxin B in microglia cells. *Journal of Pharmacological Sciences, 128*(1), 35–46.

Ghaisas, S., Maher, J., & Kanthasamy, A. (2016). Gut microbiome in health and disease: Linking the microbiome-gut-brain axis and environmental factors in the pathogenesis of systemic and neurodegenerative diseases. *Pharmacology & Therapeutics, 158*, 52–62.

Grishin, A., Bowling, J., Bell, B., Wang, J., & Ford, H. R. (2016). Roles of nitric oxide and intestinal microbiota in the pathogenesis of necrotizing enterocolitis. *Journal of Pediatric Surgery, 51*(1), 13–17.

Gubiani, D., Petrič, I., Fabbretti, E., & Urbančič, T. (2015). Mining scientific literature about ageing to support better understanding and treatment of degenerative diseases. In *Proceedings of conference on data mining and data warehouses (SIKDD 2015) at the 18th international multiconference information society (IS 2015)* (pp. 475–486).

He, C., Parra, D., & Verbert, K. (2016). Interactive recommender systems: A survey of the state of the art and future research challenges and opportunities. *Expert Systems with Applications, 56*, 9–27.

Hess, D. T., Matsumoto, A., Kim, S. O., Marshall, H. E., & Stamler, J. S. (2005). Protein S-nitrosylation: Purview and parameters. *Nature Reviews Molecular Cell Biology, 6*, 150–166.

Horn, T. F. W., Wolf, G., Duffy, S., Weiss, S., Keilhoff, G., & MacVipar, B. A. (2002). Nitric oxide promotes intracellular calcium release from mitochondria in striatal neurons. *The FASEB Journal, 16*(12), 1611–1622.

Hristovski, D., Friedman, C., Rindflesch, T. C., & Peterlin, B. (2008). Literature-based knowledge discovery using natural language processing. In *Literature-based discovery* (pp. 133–152). 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.

Jain, A. K., Murty, M. N., & Flynn, P. J. (1999). Data clustering: A review. *ACM Computing Surveys, 31*(3), 264–323.

Jensen, L. J., Saric, J., & Bork, P. (2006). Literature mining for the biologist: From information retrieval to biological discovery. *Nature Reviews Genetics, 7*, 119–129.

Jeong, J. J., Woo, J. Y., Ahn, Y. T., Shim, J. H., Huh, C. S., Im, S. H., et al. (2015). The probiotic mixture IRT5 ameliorates age-dependent colitis in rats. *International Immunopharmacology, 26*(21), 416–422.

Juršič, M., Cestnik, B., Urbančič, T., & Lavrač, N. (2012a). Bisociative knowledge discovery: An introduction to concept, algorithms, tools, and applications. In M. R. Berthold (Ed.). In *Lecture notes in artificial intelligence: 7250* (pp. 338–358). Springer.

Juršič, M., Cestnik, B., Urbančič, T., & Lavrač, N. (2012b). Cross-domain literature mining: Finding bridging concepts with CrossBee. In M. L. Maher, K. J. Hammond, A. Pease, R. P. Pérez, D. Ventura, & G. A. Wiggins (Eds.), *Proceedings of the 3rd international conference on computational creativity* (pp. 33–40). computationalcreativity.net. http://crossbee.ijs.si/. Accessed 07.05.17.

Katusic, Z. S., & Austin, S. A. (2014). Endothelial nitric oxide: Protector of a healthy mind. *European Heart Journal, 35*(14), 888–894.

Kolios, G., Rooney, N., Murphy, C. T., Robertson, D. A., & Westwick, J. (1998). Expression of inducible nitric oxide synthase activity in human colon epithelial cells: Modulation by T lymphocyte derived cytokines. *Gut, 43*, 56–63.

Kumar, V. D., & Tipney, H. J. (2014). *Biomedical literature mining*. Springer.

Langley, P. (2000). The computational support of scientific discovery. *International Journal of Human-Computer Studies, 53*(3), 393–410.

Lundberg, J. O. N., Lundbergand, J., Alving, K., & Hellstrom, P. M. (1994). Greatly increased luminal nitric oxide in ulcerative colitis. *The Lancet, 344*, 1673–1674.

Mancuso, C., & Santangelo, R. (2014). Ferulic acid: Pharmacological and toxicological aspects. *Food and Chemical Toxicology, 65*, 185–195.

Middleton, S. J., Shorthouse, M., & Hunter, J. O. (1993). Increased nitric oxide synthesis in ulcerative colitis. *The Lancet, 341*, 465–466.

Moore, D. S., McCabe, G. P., & Craig, B. A. (2007). *Introduction to the practice of statistics* ((6th ed)). W. H. Freeman and Company.

Oh, J. H., & Deasy, J. O. (2016). A literature mining-based approach for identification of cellular pathways associated with chemoresistance in cancer. *Briefings in Bioinformatics, 17*(3), 468–478.

Perovšek, M., Juršič, M., Cestnik, B., & Lavrač, N. (2016a). Empowering bridging term discovery for cross-domain literature mining in the TextFlows platform. In A. Holzinger (Ed.), *Machine learning for health informatics*. In *Lecture notes in computer science: 9605* (pp. 59–98). Springer.

Perovšek, M., Kranjc, J., Erjavec, T., Cestnik, B., & Lavrač, N. (2016b). TextFlows: A visual programming platform for text mining and natural language processing. *Science of Computer Programming: Special Issue on Knowledge-based Software Engineering, 121*, 128–152. http://textflows.org. Accessed 07.05.17.

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

Petrič, I., Urbančič, T., & Cestnik, B. (2006). Comparison of ontologies built on titles, abstracts and entire texts of articles. In *Proceedings of the 9th international multiconference information society (IS 2006)* (pp. 227–230).

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.

Phillips, R. J., & Powley, T. L. (2007). Innervation of the gastrointestinal tract: Patterns of aging. *Autonomic Neuroscience: Basic and Clinical, 136*(1–2), 1–19.

Rajpal, D. K., Qu, X. A., Freudenberg, J. M., & Kumar, V. (2014). Mining emerging biomedical literature for understanding disease associations in drug discovery. In *Biomedical literature mining*. In *Methods in molecular biology: 1159* (pp. 171–206).

Rannikko, E. H., Weber, S. S., & Kahle, P. J. (2015). Exogenous α-synuclein induces toll-like receptor 4 dependent inflammatory responses in astrocytes. *BMC Neuroscience, 16*(1), 57.

Rivera, L. R., Poole, D. P., Thacker, M., & Furness, J. B. (2011). The involvement of nitric oxide synthase neurons in enteric neuropathies. *Neurogastroenterology & Motility, 23*(11), 980–988.

Sluban, B., Juršič, M., Cestnik, B., & Lavrač, N. (2012). Exploring the power of outliers for crossdomain literature mining. In M. R. Berthold (Ed.), *Bisociative knowledge discovery: An introduction to concept, algorithms, tools, and applications*. In *Lecture notes in artificial intelligence: 7250* (pp. 325–337).

Smalheiser, N. R., & 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.

Smalheiser, N. R., & Torvik, V. I. (2008). The place of literature-based discovery in contemporary scientific practice. In P. Bruza, & M. Weeber (Eds.), *Literature-based discovery*. In *Information science and knowledge management: 15* (pp. 13–22). Springer.

Steinert, J. R., Chernova, T., & Forsythe, I. D. (2010). Nitric oxide signaling in brain function, dysfunction, and dementia. *The Neuroscientist, 16*(4), 435–452.

Swanson, D. R. (1986). Undiscovered public knowledge. *Library Quarterly, 56*(2), 103–118.

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), 29–37.

Swanson, D. R. (2008). Literature-based discovery? The very idea. In P. Bruza, & M. Weeber (Eds.), *Literature-based discovery*. In *Information science and knowledge management: 15* (pp. 3–11). Springer.

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.

Toda, N., & Okamura, T. (2012). Cerebral blood flow regulation by nitric oxide in Alzheimer's disease. *Journal of Alzheimer's Disease, 32*(3), 569–578.

Weeber, M. (2007). Drug discovery as an example of literature-based discovery. In S. Džeroski, & L. Todorovski (Eds.), *Computational Discovery of Scientific Knowledge* (pp. 290–306). Springer-Verlag.

Weeber, M., Klein, H., de Jong-van den Berg, L. T. W., & Vos, R. (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.

Xiao, J., Li, S., Sui, Y., Wu, Q., Li, X., Xie, B., et al. (2014). Lactobacillus casei-01 facilitates the ameliorative effects of proanthocyanidins extracted from lotus seedpod on learning and memory impairment in scopolamine-induced amnesia mice. *PLoS One, 9*(11), e112773.

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

Zhang, Y., Sarkar, I. N., & Chen, E. S. (2014). PubMedMiner: Mining and visualizing MeSH-based associations in PubMed. In *Proceedings of the annual symposium AMIA* (pp. 1990–1999).