GENOMICS AND COMPUTATIONAL BIOLOGY Vol. 3 No. 3 (2017): e58

Research Article

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

Bojan Cestnik[1,3,*], Elsa Fabbretti[2], Donatella Gubiani[2], Tanja Urbančič[1,2], Nada Lavrač [1,2]

[1]Jožef Stefan Institute, Jamova 39, Ljubljana, Slovenia
[2]University of Nova Gorica, Vipavska 13, Nova Gorica, Slovenia
[3]Temida d.o.o., Dunajska 51, Ljubljana, Slovenia

*Correspondence: Tel: +386 1 236 3351; Fax: +386 1 236 3356; Email: bojan.cestnik@temida.si

Received 2016-10-30; Accepted 2017-03-30

ABSTRACT

Literature-based discovery tools have been often used to overcome the problem of fragmentation of science and to assist researchers in their process of cross-domain knowledge discovery. In this paper we propose a methodology for cross-domain literature-based discovery that focuses on outlier documents to reduce the search space of potential cross-domain links and to improve search efficiency. In a previous study, literature mining tools OntoGen for document clustering and CrossBee for cross-domain bridging term exploration were combined to search for hidden relations in scientific papers from two different domains of interest, where the utility of the approach was demonstrated in a study involving PubMed papers about Alzheimer's disease and gut microbiome. This paper extends the approach by proposing a methodology, implemented as a repeatable workflow in a web-based text mining platform TextFlows, which enables easy access and execution of the methodology for the interested researcher.

KEYWORDS

Literature-based discovery, cross-domain discovery, outliers, workflows, gut microbiome, Alzheimer Disease

AVAILABILITY AND REQUIREMENTS

- Project name: LBD workflows for outlier detection
- Web: http://crossbee.ijs.si/; http://textflows.org/workflow/1844/; http://ontogen.ijs.si
- Systems: web browser; MS Windows
- Programming language: html 5, javascript
- License: GNU GPL v3

INTRODUCTION

Literature-based discovery (LBD) is an IT technology for examining hidden relations among pieces of information published in diverse and rapidly growing scientific literature. It has proved to be useful for overcoming the problem of fragmentation of science and for assisting researchers in their process of cross-domain knowledge discovery [1]. The field of LBD started to evolve with the early work of Swanson [2] and Smalheiser [3], who developed early approaches to assist the user in detecting interesting cross-domain bridging terms with a goal to discover unknown relations between previously unrelated concepts in two different domains (two corpora of medical articles) of interest. Their 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 other researchers to develop it further [1].

Several IT tools have been developed in the field of LBD, with notable results like early ARROWSMITH [3] and its extensions [4, 5], LitLinker [6], and BITOLA [7]. An example tool that we have developed in previous research is CrossBee [8], a web-based tool for bridging term (b-term) discovery and exploration, which implements an ensemble based term ranking approach to finding new connections between two predefined domains, represented by two user defined sets of biomedical articles. The research conducted by Petrič et al. [9] and Sluban et al. [10] complements this research by showing that bridging terms are substantially more frequent in documents that are outlier documents of their own domain, compared to their frequency in normal, non-outlier documents. Analogously to statistics, where an outlier is defined as an observation that falls outside the overall pattern of a distribution [11], an outlier document in the field of LBD is a document that lies outside the main group of documents of its own domain and is, therefore, in two domain settings more similar to the documents of the other explored domain than to the documents of the domain of its origin.

Recently, an online web-based platform TextFlows [12] has been developed to facilitate handy and repeatable experimentation with complex text mining tools. Using TextFlows, the interested user can construct text mining workflows composed of numerous predefined text processing components. The constructed workflows can be executed and the results displayed for further inspection and analysis. Several other web-based tools for the analysis of biomedical literature that focus on biomedical entities such as disease, drugs, genes, proteins and organs are described by Holzinger et al. [13].

Such IT tools can offer help also in integrating

data and knowledge from different contexts (e.g., biomedicine, microbiology, nutrition, etc.) to provide guidelines for evidence based interdisciplinary biomedical and clinical research. A prominent challenge of this kind is human population ageing. Average life expectancy continues to rise, causing severe demographic changes with dramatic consequences, especially regarding deteriorated health of aged individuals and the lack of resources for effectively managing ageing-related diseases. Our preliminary literature mining investigation in this field of age-related pathologies, such as neurodegenerative diseases, suggested links between dietary issues and Alzheimer's disease to be further investigated [14]. Some recent medical studies also indicate exploring connections between the digestive system, gut microbiome and neurodegenerative diseases (like Alzheimer's disease) as a promising area of biomedical research [15]. A growing number of recently published research papers in this area is a strong motivation for using text and literature mining methods to research the hypothesis that, besides causing gut problems, an imbalance of gut microbiome can be associated with memory and cognition dysfunction and brain diseases.

In our previous case study, in which we investigated potential links between gut microflora and Alzheimer's disease, the methodology relied on using literature mining tools OntoGen [16] and CrossBee [17] to search for hidden relations in the published scientific papers [18]. This paper presents an extended methodology, together with its implementation as a workflow in web-based text mining platform TextFlows [12], enabling its wider use and re-use by other researchers.

The methodology combines the process of detecting outlier documents and the literature-based discovery process, aiming to help the expert in finding implicit relationships among concepts of two different domains of interest. The underlying assumption is that while the majority of articles in the given scientific domain describe matters related to a common understanding of the domain or more intensively investigated issues, 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 [18].

This paper is organized as follows. The Methods section introduces and describes the two methods that are used for knowledge discovery: outlier document detection and closed discovery literature mining. Then, the combined methodology is presented as a two-step process, combining outlier document detection and cross-domain term exploration using the CrossBee tool, which has been implemented as a workflow in web-based platform TextFlows. The Results section illustrates the application of the methodology to Alzheimer's disease and gut microbiome domains. The Discussion section evaluates and compares the obtained results with other research, providing a summary and directions for further work.

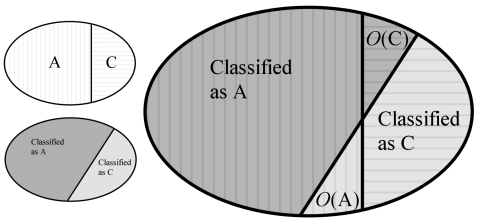

Figure 1: Detecting outlier documents by using document classification from two domains. Top-left figure presents the original document sets, bottom left the document sets as labeled by the classifier, while the outlier document sets are shown in the figure at the right.

METHODS

This section describes the background technologies that are used in our approach to cross-domain literature-based discovery. It first outlines some approaches to outlier document detection, followed by the presentation of the closed discovery process used in literature-based discovery. The third subsection describes a methodology that combines detection of outlier documents and literature based discovery with the aim of reducing the search space for potential linking terms. In the fourth subsection the implementation of the novel methodology as a workflow in TextFlows is presented.

Outlier document detection

One of the techniques that can be used to detect outlier documents is by using classification algorithms [10]. The technique works as follows. Having documents from two domains of interest we first train a classification model that distinguishes between the documents of these domains. Using the constructed model we classify all the documents. The documents that are misclassified, i.e. classified as belonging to the other domain than to the domain of its 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 to 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 vice versa. We denote the two sets of domain outlier documents with O(A) and O(C), respectively. Figure 1 illustrates this principle.

It can be shown that the majority of bridging terms can be found in outlier documents [10], as show on the gold standard migraine-magnesium domain pair, for which a confirmed list of concept bridging terms (b-terms) was made available. 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. Hence the effort needed for finding cross-domain links is substantially

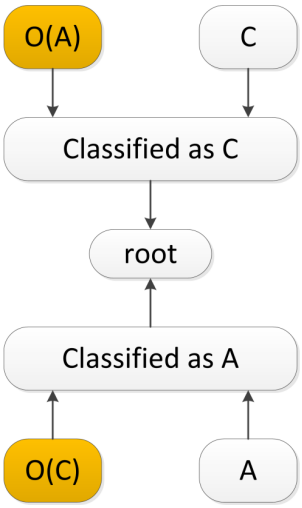

Figure 2: Documents from literatures A and C, clustered according to the OntoGen's two step approach, first using unsupervised and then supervised clustering to obtain outlier documents O(A) and O(C) of literatures A and C, respectively.

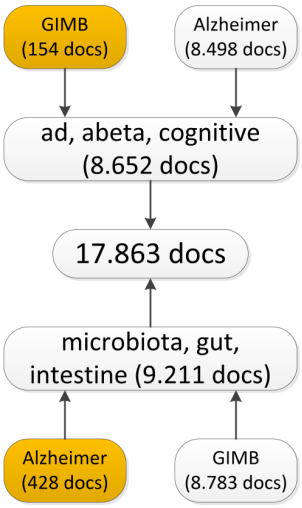

Figure 3: A cluster hierarchy constructed from the dataset of 17,863 papers with OntoGen. Two first-level clusters are labeled with extracted keywords "ad, abeta, cognitive" and "microbiota, gut, intestine". Four second level sub-clusters separate documents according to their original search keyword. Clusters containing outlier documents are shown in orange.

reduced, as it requires exploring a much smaller subset of documents, where a great majority of b-terms are present and more frequent. The achieved search space reduction has two important consequences. First, the executions of the used LBD tools are substantially faster on smaller document sets. In addition, due to specific current functional limitations of each tool, the original large sets of documents might exceed the tool's capacity. Second, faster execution times and smaller result sets containing potential bridging terms impose less strain on the involved experts, thus making their involvement in the process more efficient and effective.

An alternative approach to outlier document detection is by using document clustering. Document clustering refers to clustering methods [19], which are very popular for grouping instances in terms of their similarity, but addopted to grouping of text documents [20]. In particular, we used the OntoGen document clustering tool [16] to find outlier documents. In the approach, proposed by Petrič et al. [9], we concentrated on a specific type of outliers – the domain outliers – i.e. the documents that tend to be more similar to the documents of the opposite domain than to those of their own domain. The approach consists of two steps. In the first step, the OntoGen clustering algorithm [16] is applied to cluster the merged document set A∪C, consisting of documents from both domains A and C. The result of unsupervised clustering is a set of two document clusters: A' (i.e. a set of documents from A∪C classified as A), and C' (i.e. a set of documents from A∪C classified as C). Then, in the second step, 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, as illustrated in Figure 2.

The OntoGen tool can then be used to build two document clusters, A' and C' (where A'∪C' = AC) in an unsupervised manner, using OntoGen's 2-means clustering algorithm (see Figure 2). Cluster A' (labeled Classified as A in Figure 2) consists mainly of documents from A, but may contain also some documents from C. Similarly, cluster C' (labeled Classified as C in Figure 2) consists mainly of documents from C, but may contain also some documents from A.

Results obtained by Sluban et al. [10] and by Petrič et al. [9] confirm the hypothesis that most bridging terms appear in outlier documents and that by considering only outlier documents the search space for b-term identification can be largely reduced. In this way, we can substantially reduce the search space for finding b-term candidates.

Closed discovery with literature mining

Early work in literature-based discovery (LBD) by Swanson [2] and Smalheiser et al. [3] resulted in an approach to assist the user by detecting interesting cross-domain terms with a goal to uncover the possible relations between previously unrelated concepts. The online ARROWSMITH system, developed by Smalheiser et al. [3], takes as input two sets of titles of scientific papers from disjoint domains (disjoint document corpora) A and C, and lists terms that are common to A and C; the resulting bridging terms b are further investigated by the user for their potential to generate new scientific hypotheses. Their approach, known as the ABC model of knowledge discovery (note that 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), addresses several settings, including the closed discovery setting, introduced by Weeber et al. [21], where two initially separate domains A and C are specified by the user at the beginning of the discovery process, and the goal is to search for bridging concepts (terms) b in B in order to support the validation of the hypothesized connection

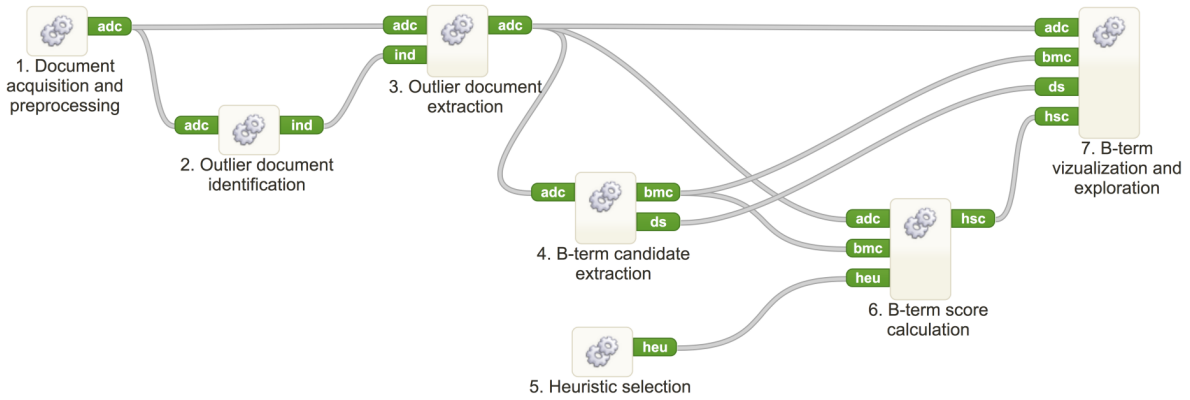

Figure 4: A top-level workflow of the proposed methodology in TextFlows [12]. The acronyms of component outputs and inputs denote connector's data type and are explained in Abbreviations section.

between A and C.

By studying two separate literatures, the literature on migraine headache and the articles on magnesium, Swanson [22] discovered connections supportive for the hypothesis that magnesium deficiency might cause migraine headache. This well-known example has become the gold standard in the literature mining field and has been used as a benchmark in several studies, including our own work, developed by Juršič et al. [8], Sluban et al. [10] and Petrič et al. [9], which is the basis for the methodology presented in this paper.

Estimating which of the terms have a high potential for interesting discoveries is a challenging research question. Juršič et al. [8] suggested a solution in which candidate bridging terms are ranked by ensemble voting of heuristics. The methodology was implemented in the CrossBee system, an off-the-shelf solution for finding bridges between two user-defined domains/literatures. Supplementary functionalities and visualizations make CrossBee a user-friendly web application, helping the experts to efficiently investigate cross-domain links.

Instead of a single step outlier detection process, used in Petrič et al. [9], we here use a two-step outlier detection process using OntoGen, illustrated in Figure 3 when applying OntoGen on the actual documents of this application. The method uses domains A and C, and builds a joint document set AC (i.e. A∪C). For this intention, two individual sets of documents (e.g. titles, abstracts or full texts of scientific articles), one for each domain under research (namely, literature A and literature C), are automatically retrieved from bibliographic databases or extracted from other document sources. We consider all the terms and not just the medical ones. A list of 523 English stop words is then used to filter out meaningless words, and English Porter stemming is applied.

Combined methodology proposed in our previous work

The particular methodology applied in this work follows our previous work in outlier detection [9] using a document clustering and exploration tool OntoGen [16]. In this methodology, each document from the two literatures is an instance, represented by a set of words using frequency statistics based on the Bag of Words (BoW) text representation [23]. The BoW vector enables to measure content similarity of documents. Content similarity computation is performed with OntoGen, which was designed for interactive data-driven construction of topic ontologies [16]. Content similarity is measured by cosine distance and the standard TF*IDF (term frequency inverse document frequency) word weighting measure [24], where high frequency of co-occurring words in documents indicates high document similarity.

The cosine similarity measure, commonly used in information retrieval and text mining to determine the semantic closeness of two documents where document features are represented using the BoW vector space model, is used to position the documents according to their similarity to the representative document (centroid) of a selected domain. Documents positioned based on the cosine similarity measure can be visualized in OntoGen by a similarity graph with cosine similarity values that fall within the [0, 1] interval. Value 0 means extreme dissimilarity, where two documents share no common words, while value 1 represents the similarity between two semantically identical documents in the BoW representation.

New methodology and its implementation as a repeatable workflow in TextFlows

Compared to this early approach to outlier document detection using OntoGen, an upgraded method is proposed in this paper. We implemented the presented approach in a web-based[1] text mining platform called TextFlows [12] that is used to construct and execute advanced text mining workflows. The workflow (see Figure 4) consists of seven steps implemented as sub-processes. The connections between sub-processes represent the flow of documents from one sub-process to another. In overview, steps 1–3 represent Outlier detection part, and steps 4–7

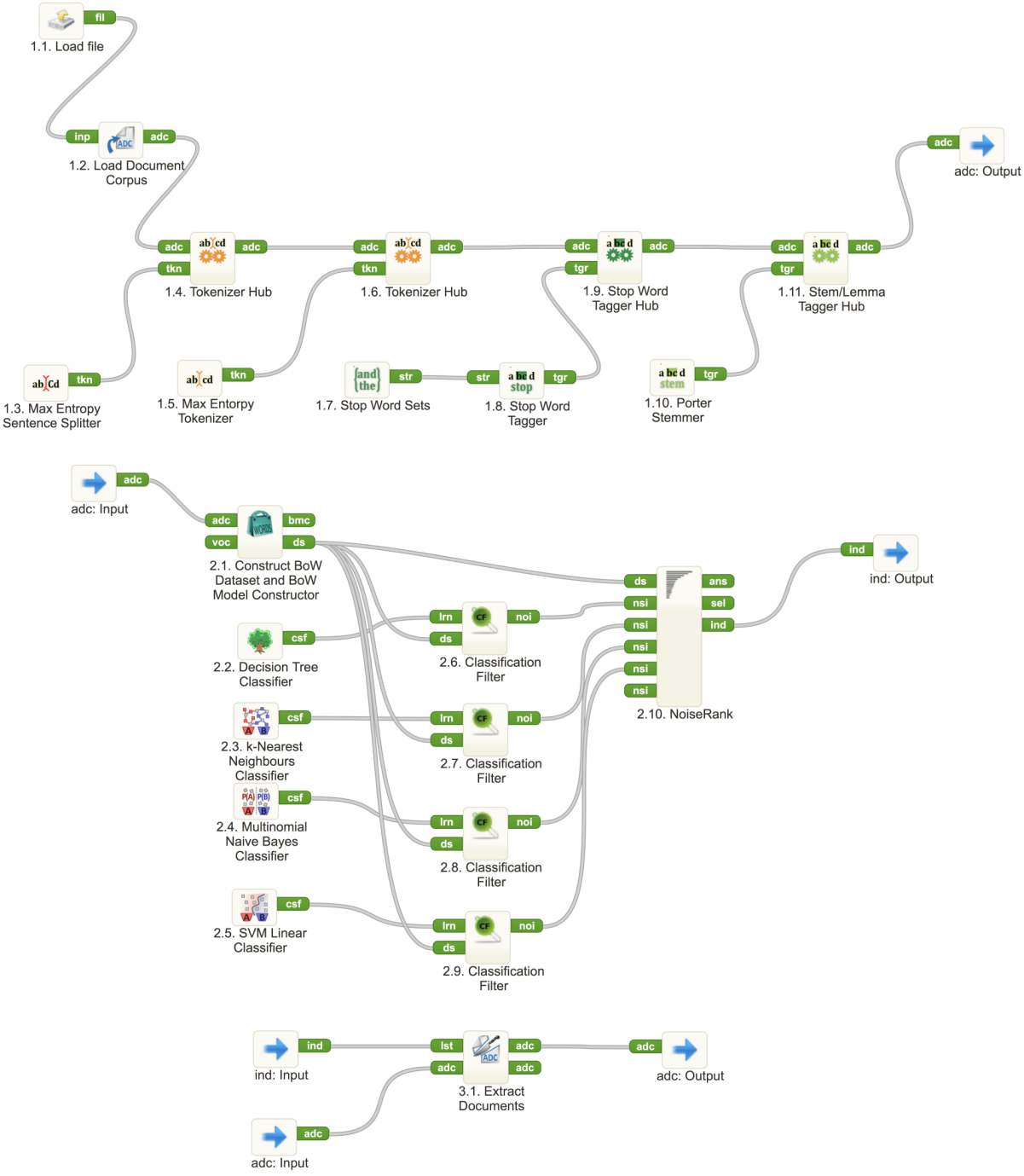

Figure 5: Detailed workflows for the first three sub-processes that are used to identify and extract outlier documents.

represent Cross-domain exploration part. Note that in the second part the role of the domain expert is crucial. Each sub-process is further presented and explained in Figures 5 and 7.

In the first three steps of the workflow depicted in Figure 4 the outlier documents are identified and extracted (instead of OntoGen we use NoiseRank [25] as implemented in TextFlows). The goal of this phase is to extract a set of outlier documents from the whole corpus of input documents. Consequently, by decreasing the size of the input set of documents the second phase becomes more focused, efficient and effective.

In the last four steps of the workflow in Figure 4 components that constitute CrossBee [17] are executed to conduct expert-guided b-term analysis. Here, the goal is to further prepare the input documents for b-term visualization and exploration.

In Figure 5 the detailed workflows for the first three sub-processes 1. Document acquisition and preprocessing, 2. Outlier document identification and 3. Outlier document extraction are presented.

The main task of the first sub-process, labeled 1. Document acquisition and preprocessing, is to read the input set of documents from a file and to transform text documents into a predefined well-structured data representation containing a set of relevant features for further processing [20]. Each document is represented by a BoW vector of numerical values [23], one for each feature of the selected representational model. In TextFlows, the preprocessing techniques are based on standard text mining concepts [1] and include Tokenization, Stop words tagging, and Stem/Lemma tagging, which are all present in the workflow of the first sub-process. Note that the applied preprocessing techniques do not include more complex text mining techniques like Part of speech tagging, Entity detection and Relation detection [20].

The purpose of the two workflows of the second and third sub-processes in Figure 5 is to identify and extract outlier documents from the initial document corpus. In contrast to the approach for outlier detection with OntoGen, described in Methods section, NoiseRank component implements a different strategy [25]. Here, classifiers are used to detect atypical documents in categorized document corpora, which can be considered as outliers of their own document category. In the workflow shown in Figure 5 we used four classifiers implemented in TextFlows to determine outlier documents: 2.2. Decision Tree classifier, 2.3. k-Nearest Neighbours classifier, 2.4. Multinomial Naïve Bayes classifier, and 2.5. SVM Linear classifier [20].

The main purpose of NoiseRank component as implemented in TextFlows (widget 2.10. in Figure 5) is to support domain experts in identifying noisy, outlier or erroneous data instances [25]. The users are able to select the noise detection algorithms to be used in the ensemble-based noise detection process. The NoiseRank methodology workflow returns a visual representation of a list of potential outlier documents, ranked according to the decreasing number of noise detection algorithms which identified a document as outlier. So, in addition, the users can obtain a visual representation of top-ranked outlier documents as shown in Figure 6. The feature allows also for manual intervention (inclusion or exclusion from the list of outliers). Note that in our experiments we have chosen the default select-all option.

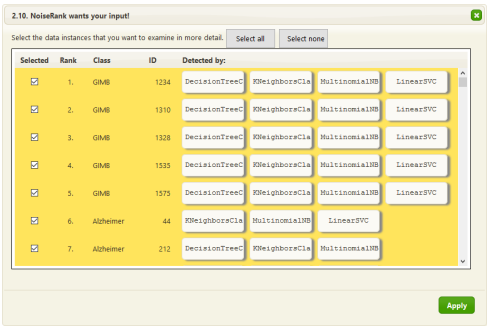

Figure 6: User interface for NoiseRank widget for inspecting and selecting outlier documents in TextFlows component.

In Figure 7 the detailed workflows for sub-processes 4. B-term candidate extraction, 5. Heuristic selection, 6. B-term score calcualtion and 7. B-term vizualization and exploration are presented.

The component 4.1 first constructs BoW representation of the remaining outlier documents. Then, in sub-process 5., the heuristics used to evaluate b-term potential are selected [8]. Sub-process 6. consists of a single widget labeled 6.1. Calculate Term Heuristic Scores. This widget takes as an input several heuristics specifications and performs the actual calculations of the heuristics on each and every b-term. 7.4. Explore in CrossBee widget, which exports the final ranking results and the annotated document corpus into web application CrossBee, is the most important part of sub-process 7. This component enables manual exploration of potential b-terms and respective documents.

The proposed methodology of using outlier documents to speed-up literature based discovery process was first implemented as a combination of two text mining tools: OntoGen and CrossBee [18]. In this paper we extended the methodology by presenting it as a workflow in TextFlows platform [12]. In principle, the two implementations share the same philosophy; however, they are not identical because they use different methods for detecting outlier documents.

OntoGen is a desktop application freely available for download[2] that runs on Windows operating system. As mentioned earlier, CrossBee[3] and TextFlows are both web applications accessible through a web browser. The CrossBee part of the methodology corresponds to sub-processes 4–7 in TextFlows workflow in Figure 4. Since the CrossBee components are also implemented in TextFlows, the two parts are virtually identical. However, the outlier detection part (sub-processes 1–3 in Figure 4) differs between the two approaches. While OntoGen is efficient and can process several thousands of documents in a couple of minutes, it might take TextFlows workflow component several hours to process

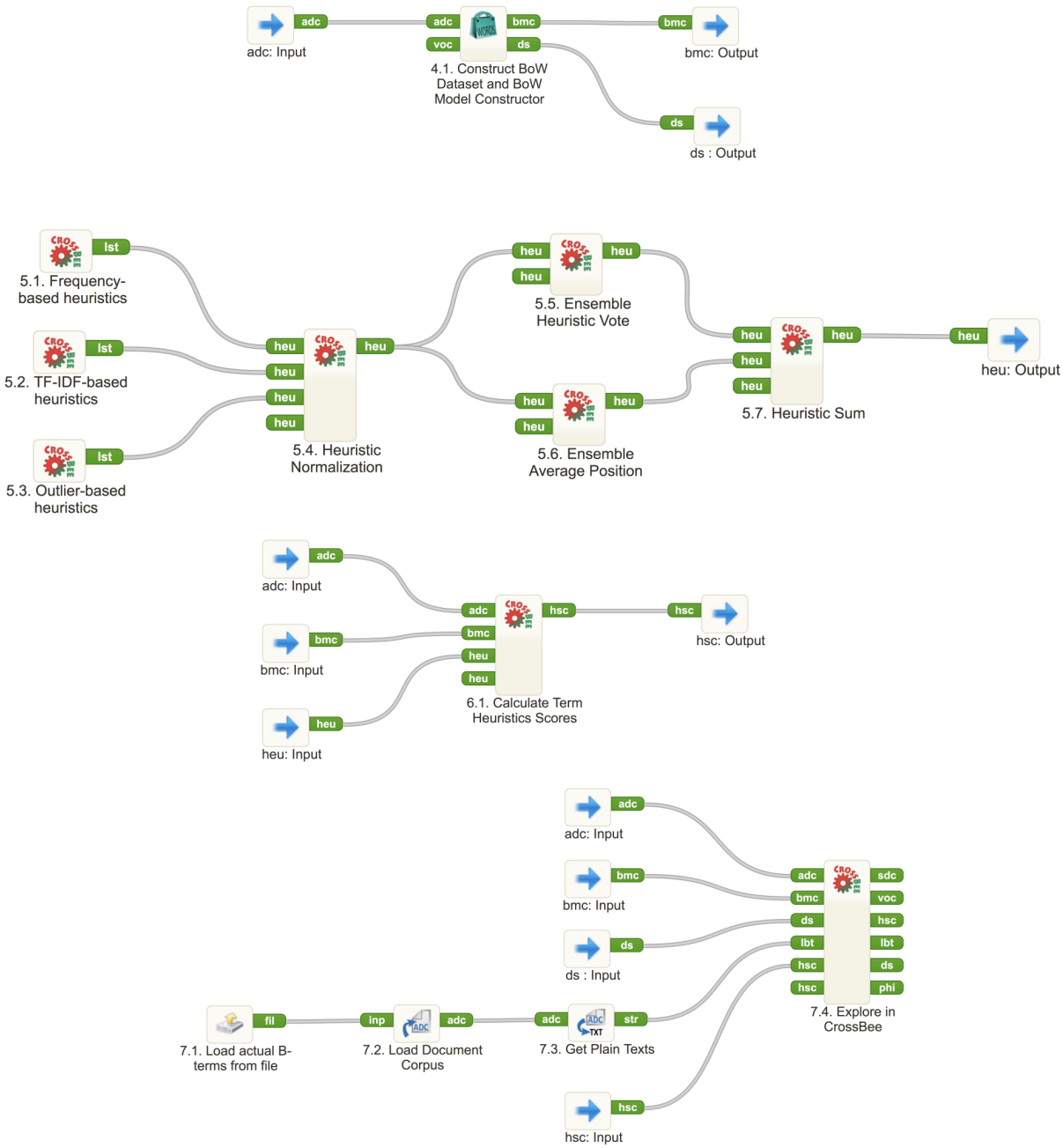

Figure 7: Workflows for the last four sub-processes that are implemented by using CrossBee components in TextFlows.

the same workload. On the other hand, TextFlows workflows are more transparent and allow for easier parametrization and user intervention. So, for efficiency reasons, the outlier detection part in our experiment was carried out in OntoGen.

RESULTS

In the practical experiment with the proposed methodology we have demonstrated the utility of outlier detection to finding links between the literatures from two distinctive domains: gut microbiome and Alzheimer's disease. In the first experiment, we downloaded two input sets of documents from PubMed: 83,322 papers obtained from query "Alzheimer" and 73,960 papers obtained from query "(gut OR intestinal) AND (microbiota OR bacteria)". We used titles and abstracts based on previous experimental evidences described in [26]. Due to current functional limitations of the tool, these sets were reduced by eliminating documents older than two years (i.e. we included years 2014 and 2015) and documents with incomplete title or abstract. As a result, we obtained 8,934 "Alzheimer" papers and 8,937 "gut microbiota" papers. Note that in further work we consider experimenting with more elaborated search criteria using MeSH terms filtering, which would result in a reduced set of outlier documents that needed to be explored by the expert. Related to the workflow implementation in Figure 5, which is included as an illustration of the methodology, we, for the proof of concept, extracted a random sample of 1.971 papers from the whole set of 17.863 papers (11%) and obtained 248 outlier documents for further processing. It took the first three sub-processes 7 minutes to complete the outlier detection phase.

On the joint set of 17,863 papers we constructed a two-level document hierarchy with OntoGen, following the approach to identify outlier documents, proposed in [9]. At the first level, after transforming the documents into a feature vector format, the documents were clustered according to the cosine similarity into two distinct document clusters. At the second level, as shown in Figure 3, each of the two clusters was further separated according to the document search origin ("Alzheimer" or "gut microbiota") into two sub-clusters.

Based on the constructed cluster hierarchy from Figure 3 we obtained 582 outlier documents (428 from "Alzheimer" and 154 from "gut microbiota"). We further explored these documents in the on-line Cross-domain Bisociative Exploration Tool CrossBee. By processing the documents from the two separate domains (literatures) of interest, CrossBee extracted 4.723 terms as potential terms connecting the two domains. The terms in CrossBee were ranked according the estimated bridging term potential [17], tending to push more interesting terms to the top of the list. Even though the list of potential bridging terms is ordered according to the term's potential, browsing and analyzing the terms from the list still presents a substantial burden for the domain expert and supportive team.

To further reduce the size of the potential bridging term list, the domain expert prepared a special list of 289 potentially interesting terms. They were extracted from 42 papers obtained from PubMed search query "gut+Alzheimer". Common terms (such as dementia, ageing, neurodegeneration, inflammation, microbiome, probiotics etc.) and specific molecular factors and pathways (such as the names of transcription factors, trophic factors, proteins misfolded or aggregated, mechanisms of oxidative stress and lipid synthesis, etc.), were manually identified in title, abstracts, and keywords of such publications. 55 terms from this list appeared also among the 4.723 terms extracted by CrossBee.

Several terms from the prepared list were also suggested by CrossBee and identified by the expert as prospective for further exploration. The candidate terms include, for example, "cox" (ranked 80 in CrossBee), "mucosa" (122), "beta amyloid" (749), "nitric oxide synthase" (1153) and "phenylalanine" (2080). In the evaluation and discussion that followed the relevant papers for each b-term candidate were reviewed and searched for potential clues justifying further investigation.

Here we give an example how a promising b-term identified by the expert can be further explored using literature knowledge [18]. The b-term 'nitric oxide synthase' (i.e. NOS) [27] can be interpreted in view of microbiome contribution in NOS-mediated inflammation at gut level [28–30] [28-30]. Although not yet demonstrated, it is likely that such effects dysregulate the brain-gut communication. In pathological conditions, such as during inflammation or in the presence of environmental stressors or ageing, abnormal iNOS function results in oxidative effects and neurodegeneration. In particular, in the gut, iNOS induces intestinal barrier damage [31], and in the brain causes nitrosylation of proteins and cell death with neurological consequences like dementia, Alzheimer or Parkinson diseases [32, 33]. Even though NO is locally produced, its effects can diffuse systemically, thus influencing the progression of the disease [34–36] [34-36]. Bioactive nutrients possibly modulate individual microbiome responses limiting inflammation and free radical synthesis, including NO [37]. Although further studies are necessary to clarify the consequences of pathological NO signaling in different tissues, the study of NO/iNOS-targeted therapeutic strategies might have a clinical benefit [38]. Our study suggests that such a validated IT process identify new common molecular targets that can be used as multi-purpose drugs, able to block multiple processes such as oxidative and inflammatory effects with high neuroprotection relevance for peripheral and central neurons. Note that the combined query "Alzheimer gut nitric oxide" in PubMed revealed that there are no articles that correspond to this query, indicating that this connection was not previously explored in the literature available in PubMed.

DISCUSSION

The presented methodology integrates the process of exploring outlier documents and the literature-based discovery process. The methodology was first executed by combining two text mining tools OntoGen [16] and CrossBee [17]. Apart from constructing the document clustering from the input set of documents and finding

the keywords describing the two document classes, OntoGen can be used to narrow down the search for bridging terms in CrossBee by identifying the set of outlier documents. In such way, the search in CrossBee can be more focused, efficient and effective. This opens the door for many new applications in which the size of the search space is a severe limiting factor.

We have also implemented the proposed methodology as a workflow in TextFlows platform [12]. Besides a clear demonstration and explanation of the procedural steps required by the methodlogy, it enables easy access and execution of the methodology for the interested researcher to experiment with various input files and parameter settings.

In our experiment it turned out that the expert from the biomedical field played a crucial role in the document exploration process, as the interdisciplinary collaboration between the team members allowed for efficient investigation of bridging terms suggested by CrossBee. Among them, in our first experiment, we identified and decided to focus on the term "nitric oxide synthase" [27] as a promising novel bridging term. From the set of articles that were used as input to CrossBee the expert carefully reviewed the four articles containing this term; three from the "Alzheimer" domain [34–36] [34-36] and one from the "gut microbiota" domain [30]. In addition, the combined query "Alzheimer gut nitric oxide" in PubMed revealed that there are no articles that correspond to this query, which is especially interesting, indicating that this connection was not previously explored in the available literature.

Due to the rapid growth of scientific literature and the subsequent huge search space of possibilities, IT offers a useful support to solve complex interdisciplinary questions in finding cross-domain links leading to new insights and discoveries. In this paper we suggest a methodology that combines two possible approaches to overcome this problem. First it identifies the parts of the search space with increased probability of finding good candidate terms/concepts thus restricting the huge amounts of existing literature to more manageable sources to be explored, and then it estimates the potential of candidate links for new discoveries, enabling the user to concentrate on the most promising ones.

ACKNOWLEDGEMENTS

The authors wish to thank Matic Perovšek for his work on text mining workflow platform TextFlows, Matjaž Juršič for his work on CrossBee, Borut Sluban and Ingrid Petrič for contributing to the research on outlier document detection. This study was partially funded by the Slovenian Research Agency program Knowledge Technologies.

AUTHOR CONTRIBUTIONS

BC carried out data analysis, constructed methodlogy workflow in TextFlows and wrote the text. EF provided the list of potentially interesting bridging terms, evaluated the results of data analysis and wrote the text. DG, TU and NL carried out data analysis and wrote the text.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ABBREVIATIONS

AD: Alzheimer's disease
BoW: Bag of Words
NO: nitric oxide
NOs: nitric oxide synthase
TF*IDF: term frequency inverse document frequency
Acronyms of component data types [12]:
adc: Annotated Document Corpus
ans: All Noise
bmc: BoW Model Constructor
csf: Classifier
ds: BoW Dataset
fil: File
hsc: Heuristic Scores
heu: Heuristic Specification
ind: Selected Indices
inp: Input can be a string (str) or a file (fil)
lbt: List of Bridging Terms
lrn: Learner
lst: List of Document Indexes
noi, nsi: Noisy Instances
phi: Primary Heuristic Index
sdc: Serialized Annotated Document Corpus
sel: Selected Instances
str: Texts String with all documents in Annotated Document Corpus
tgr: Tagger (Stop words)/Stemmer
tkn: Tokenizer object and its arguments
voc: Controlled Vocabulary

REFERENCES

1. Bruza P, Weeber M. **Literature-based discovery**. vol. 15 of Springer Series in Information Science and Knowledge Management. Springer-Verlag Berlin Hiedelberg; 2008. doi:10.1007/978-3-540-68690-3.
2. Swanson DR. **Medical literature as a potential source of new knowledge**. Bulletin of the Medical Library Association. 1990;78(1):29–37.
3. Smalheiser NR, Swanson DR. **Using ARROWSMITH: a computer-assisted approach to formulating and assessing scientific hypotheses**. Comput Methods Programs Biomed. 1998;57(3):149–53.
4. Lindsay RK, Gordon MD. **Literature-based Discovery by Lexical Statistics**. Journal of the Association for Information Science and Technology. 1999;50(7):574–587. doi:10.1002/(SICI)1097-4571(1999)50:7<574::AID-ASI3>3.0.CO;2-Q.
5. Gordon MD, Lindsay RK. **Toward discovery support systems: A replication, re-examination, and extension of Swanson's work on literature-based discovery of a connection between Raynaud's and fish oil**. Journal of the American Society for Information Science. 1996;47(2):116–128. doi:10.1002/(SICI)1097-4571(199602)47:2<116::AID-ASI3>3.0.CO;2-1.
6. Yetisgen-Yildiz M, Pratt W. **Using statistical and knowledge-based approaches for literature-based discovery**. Journal of Biomedical Informatics. 2006;39(6):600–611. doi:10.1016/j.jbi.2005.11.010.
7. Hristovski D, Peterlin B, Mitchell JA, Humphrey SM. **Using literature-based discovery to identify disease candidate genes**. International Journal of Medical Informatics. 2005;74(2–4):289–298. doi:10.1016/j.ijmedinf.2004.04.024.
8. Juršič M, Cestnik B, Urbančič T, Lavrač N. **Bisociative Literature Mining by Ensemble Heuristics**. In: Bisociative Knowledge Discovery: An Introduction to Concept, Algorithms, Tools, and

Applications. Berlin, Heidelberg: Springer Berlin Heidelberg; 2012. p. 338–358. doi:10.1007/978-3-642-31830-6_24.

9. Petrič I, Cestnik B, Lavrač N, Urbančič T. **Outlier Detection in Cross-Context Link Discovery for Creative Literature Mining**. The Computer Journal. 2012;55(1):47–61. doi:10.1093/comjnl/bxq074.
10. Sluban B, Juršič M, Cestnik B, Lavrač N. **Exploring the Power of Outliers for Cross-Domain Literature Mining**. In: Bisociative Knowledge Discovery: An Introduction to Concept, Algorithms, Tools, and Applications. Berlin, Heidelberg: Springer Berlin Heidelberg; 2012. p. 325–337. doi:10.1007/978-3-642-31830-6_23.
11. Moore D, McCabe G, Craig B. **Introduction to the Practice of Statistics**. 6th ed. New York: W.H. Freeman; 2007.
12. Perovšek M, Kranjc J, Erjavec T, Cestnik B, Lavrač N. **TextFlows: A visual programming platform for text mining and natural language processing**. Science of Computer Programming. 2016;121:128–152. doi:https://doi.org/10.1016/j.scico.2016.01.001.
13. Holzinger A, Yildirim P, Geier M, Simonic KM. **Quality-Based Knowledge Discovery from Medical Text on the Web**. In: Quality Issues in the Management of Web Information. Berlin, Heidelberg: Springer Berlin Heidelberg; 2013. p. 145–158. doi:10.1007/978-3-642-37688-7_7.
14. Gubiani D, Petrič I, Fabbretti E, Urbančič T. **Mining scientific literature about ageing to support better understanding and treatment of degenerative diseases**. In: Conference on Data Mining and Data Warehouses (SiKDD 2015); 2015. .
15. Ghaisas S, Maher J, Kanthasamy A. **Gut microbiome in health and disease: Linking the microbiome–gut–brain axis and environmental factors in the pathogenesis of systemic and neurodegenerative diseases**. Pharmacology and Therapeutics. 2016;158:52–62. doi:https://doi.org/10.1016/j.pharmthera.2015.11.012.
16. Fortuna B, Grobelnik M, Mladenić D. **Semi-automatic data-driven ontology construction system**. In: Conference on Data Mining and Data Warehouses (SiKDD 2006); 2006. .
17. Juršič M, Cestnik B, Urbančič T, Lavrač N. **Cross-domain literature mining: Finding bridging concepts with CrossBee**. In: Proceedings of the 3rd International Conference on Computational Creativity; 2012. p. 33–40.
18. Cestnik B, Fabbretti E, Gubiani D, Lavrač N, Urbančič T. **Exploring outlier documents to investigate potential links between gut microbiota and Alzheimer's disease**. In: IWBBIO 2016 Internatinal Work-Conference on Bioinformatics and Biomedical Engineering, Proceedings of Extended Abstracts; 2016. p. 475–486.
19. Kaufman L, Rousseeuw PJ. **Finding groups in data: an introduction to cluster analysis**. Wiley series in probability and mathematical statistics. Hoboken, N.J.: Wiley; 2005. doi:10.1002/9780470316801.
20. Feldman R, Sanger J. **The text mining handbook : advanced approaches in analyzing unstructured data**. Cambridge ; New York: Cambridge University Press; 2007.
21. Weeber M, Klein H, de Jong-van den Berg LTW, Vos R. **Using Concepts in Literature-based Discovery: Simulating Swanson's Raynaud-fish Oil and Migraine-magnesium Discoveries**. Journal of the Association for Information Science and Technology. 2001 May;52(7):548–557. doi:10.1002/asi.1104.abs.
22. Swanson DR. **Migraine and magnesium: eleven neglected connections**. Perspectives in Biology and Medicine. 1988;31(4):526–557.
23. Sebastiani F. **Machine Learning in Automated Text Categorization**. ACM Comput Surv. 2002 Mar;34(1):1–47. doi:10.1145/505282.505283.
24. Salton G, Buckley C. **Term-weighting approaches in automatic text retrieval**. Information Processing and Management. 1988;24(5):513 – 523. doi:http://dx.doi.org/10.1016/0306-4573(88)90021-0.
25. Sluban B, Gamberger D, Lavrač N. **Ensemble-based noise detection: noise ranking and visual performance evaluation**. Data Mining and Knowledge Discovery. 2014;28(2):265–303. doi:10.1007/s10618-012-0299-1.
26. Petrič I, Urbančič T, Cestnik B. **Comparison of ontologies built on titles, abstracts and entire texts of articles**. In: Conference on Data Mining and Data Warehouses (SiKDD 2006); 2006. p. 227–230.
27. Katusic ZS, Austin SA. **Endothelial nitric oxide: protector of a healthy mind**. European Heart Journal. 2014;35(14):888. doi:10.1093/eurheartj/eht544.
28. Baruch K, Kertser A, Porat Z, Schwartz M. **Cerebral nitric oxide represses choroid plexus NFkappaB-dependent gateway activity for leukocyte trafficking**. EMBO J. 2015;34(13):1816–28. doi:10.15252/embj.201591468.
29. Derkinderen P, Rouaud T, Lebouvier T, Bruley des Varannes S, Neunlist M, De Giorgio R. **Parkinson disease: the enteric nervous system spills its guts**. Neurology. 2011;77(19):1761–1767. doi:10.1212/WNL.0b013e318236ef60.
30. Xiao J, Li S, Sui Y, Wu Q, Li X, Xie B, et al. **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. 2014;9(11):e112773. doi:10.1371/journal.pone.0112773.
31. Grishin A, Bowling J, Bell B, Wang J, Ford HR. **Roles of nitric oxide and intestinal microbiota in the pathogenesis of necrotizing enterocolitis**. Journal of Pediatric Surgery. 2016;51(1):13–17. doi:10.1016/j.jpedsurg.2015.10.006.
32. Hess DT, Matsumoto A, Kim SO, Marshall HE, Stamler JS. **Protein S-nitrosylation: purview and parameters**. Nat Rev Mol Cell Biol. 2005;6(2):150–166. doi:10.1038/nrm1569.
33. Horn TF, Wolf G, Duffy S, Weiss S, Keilhoff G, MacVicar BA. **Nitric oxide promotes intracellular calcium release from mitochondria in striatal neurons**. FASEB Journal. 2002;16(12):1611–1622. doi:10.1096/fj.02-0126com.
34. Gan P, Zhang L, Chen Y, Zhang Y, Zhang F, Zhou X, et al. **Anti-inflammatory effects of glaucocalyxin B in microglia cells**. Journal of Pharmacological Sciences. 2015;128(1):35–46. doi:10.1016/j.jphs.2015.04.005.
35. Mancuso C, Santangelo R. **Ferulic acid: pharmacological and toxicological aspects**. Food and Chemical Toxicology. 2014;65:185–195. doi:10.1016/j.fct.2013.12.024.
36. Rannikko EH, Weber SS, Kahle PJ. **Exogenous alpha-synuclein induces toll-like receptor 4 dependent inflammatory responses in astrocytes**. BMC Neuroscience. 2015;16:57. doi:10.1186/s12868-015-0192-0.
37. Jeong JJ, Woo JY, Ahn YT, Shim JH, Huh CS, Im SH, et al. **The probiotic mixture IRT5 ameliorates age-dependent colitis in rats**. Int Immunopharmacol. 2015;26(2):416–422. doi:10.1016/j.intimp.2015.04.021.
38. Broom L, Marinova-Mutafchieva L, Sadeghian M, Davis JB, Medhurst AD, Dexter DT. **Neuroprotection by the selective iNOS inhibitor GW274150 in a model of Parkinson disease**. Free Radical Biology and Medicine. 2011;50(5):633–40. doi:10.1016/j.freeradbiomed.2010.12.026.

NOTES

[1] http://textflows.org
[2] http://ontogen.ijs.si
[3] http://crossbee.ijs.si