Advance Access publication on 2 November 2010
doi:10.1093/comjnl/bxq074

Outlier Detection in Cross-Context Link Discovery for Creative Literature Mining

INGRID PETRIČ[1], BOJAN CESTNIK[2,3], NADA LAVRAČ[3,1] AND TANJA URBANČIČ[1,3,*]

[1]*University of Nova Gorica, Vipavska 13, 5000 Nova Gorica, Slovenia*
[2]*Temida, d.o.o., Dunajska 51, 1000 Ljubljana, Slovenia*
[3]*Jožef Stefan Institute, Jamova 39, 1000 Ljubljana, Slovenia*
**Corresponding author: tanja.urbancic@ung.si*

This paper investigates the role of outliers in literature-based knowledge discovery. It shows that detecting interesting outliers which appear in the literature on a given phenomenon can help the expert to find implicit relationships among concepts of different domains. 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, the exploration of outliers may lead to the detection of scientifically interesting bridging concepts among disjoint sets of scientific articles. The proposed approach contributes to cross-context link discovery by proving the utility of outlier detection for finding bisociative links in the process of autism literature exploration, as well as by uncovering implicit relationships in the articles from the migraine domain.

Keywords: outliers; bridging concepts; bisociations; literature mining; knowledge discovery

Received 25 March 2010; revised 11 September 2010
Handling editor: Einoshin Suzuki

1. INTRODUCTION

In statistics, an outlier is an observation that is numerically distant from the rest of the data, or more formally, it is an observation that lies outside the overall pattern of a distribution [1]. In many data sets outliers are due to data measurement errors, therefore it would be best to discard them from the data. However, there are also cases where outliers actually led to the discovery of intriguing new information. For instance, outlier mining has already proved to have important applications in fraud detection and network intrusion detection [2–4]. Similarly, studying outliers proved important also in economics, particularly in finance and business, where rare events can be a sign of interesting unusual activities or observations like, for instance, potential sales opportunities [5].

A specifically challenging aspect of outlier detection is emerging within climate research and extreme weather events prediction. In addition to investigating outliers as means for interesting discoveries, exploring rare concepts can also lead to important discoveries. There has been much interest in investigating the impacts, intensity and distribution of rare extreme events over a certain period of time [6]. Current attention to rare weather phenomena is driven by their possibility to become regionally more variable or extreme menace to human life, civil infrastructure and natural ecosystems, which may have significant socioeconomic impacts [7].

Rarity as a principle has been extensively researched also in the field of ecology statistics [8]. These investigations explore rarity driven by biodiversity and conservation policies [9]. Considering this fact, special concern of ecologists has been devoted to studying rare species [10]. They recognized two syndromes of rarity: habitat-limited species that are rare because their habitat was rare, and dispersal-limited species that are rare because they stayed behind due to a catastrophic disturbance-forced turnover of their habitat. While ecologists' primary concern is preventing the extinction of rare species, they also identified the potential of dispersal-limited species to adapt to the changed environment.

This paper investigates the role of outliers in the area of literature mining, contributing a novel approach to text mining [11]. It explores the utility of outliers in a nonstandard text mining task of cross-context link discovery. The motivation for our focus on outlier documents in the literature has grounds in the literature on context-crossing associations, called *bisociations*, introduced in Koestler's book *The Act of Creation* [12]. According to Koestler, a bisociation is a result of literal processes of the mind when making completely new associations between concepts from contexts (domains/categories/classes) that are usually considered separate. At the same time, Mednick's introduced *associative creativity theory* [13] which defines creative thinking as the capacity of generating new combinations of distinct associative elements (e.g. words). He explained how thinking about the concepts that are not strictly related to the elements under investigation inspires unforeseen useful connections among these elements. Consequently, exploration of bisociations may considerably improve the human creative process. Through the history of science, this mechanism has been a crucial element of progressive insights and paradigm shifts. Nevertheless, no comprehensive ICT methodology has yet been developed on this basis.

This paper provides evidence that our method of outlier document exploration can contribute to this particular approach to cross-context scientific discovery, which is based on an existing, but hitherto not computationally implemented notion of bisociation. The presented approach to creative knowledge discovery from text documents is based on exploring interesting terms in outlier documents, and is used to detect implicit relationships across different domains of expertise. It is an approach to *closed discovery* as defined by Weeber *et al.* [14], where two domains of interest are identified by the expert prior to starting the knowledge discovery process. As opposed to closed discovery, *open discovery* [14] leads the creative knowledge discovery process from a given starting domain towards a yet unknown second domain which at the end of this process turns out to be connected with the first one. Note that open discovery corresponds to hypothesis generation and closed discovery to hypothesis testing. Both ways of knowledge discovery can be supported by software tools such as RaJoLink [15].

The method proposed in this paper focuses on terms in outlier documents in the literature from two given domains/contexts (which corresponds to the closed discovery setting), with the aim of detecting bridging concepts (linking terms), enabling the exploration of the potentially interesting bisociative links between the two domains. These links may be indicative of new insights/discoveries and as such, they may support bisociative knowledge discovery.

The proposed method has been applied to the domain of autism. Autism belongs to a group of pervasive developmental disorders that are portrayed by an early delay and abnormal development of cognitive, communication and social interaction skills of a person [16]. It is a very complex and not yet sufficiently understood domain, where precise causes are still unknown; hence we have chosen it as our domain of investigation. It has also been evaluated on the domain of migraine which has been previously explored by Swanson [17].

The research presented in this paper aims at finding cross-context links between concepts from two disparate literature sources A and C, based on exploring outliers in the articles of the two domains. The main contributions of this paper are presented in Sections 2.1 and 2.3, Section 3 and most of Sections 4–6. A difference compared to our previous work [15] is that the closed discovery setting is now explored also from the point of view of bisociative cross-context link discovery (see Section 2). Moreover, we present a new methodology based on the assumption that by exploring outlier documents it is faster to discover bridging concepts that can establish previously unknown links between literatures A and C.

The paper is organized as follows. Section 2 introduces a novel correspondence between two established creative knowledge discovery frameworks: the Koestler's bisociative link discovery framework [12] and the Swanson's ABC model of closed discovery in literature mining [17]. Section 3 sets the stage for using outliers in creative knowledge discovery, connecting this work with our previous literature mining framework explored in the RaJoLink methodology [15]. Section 4 introduces outliers as means for guiding the knowledge discovery process, and identifies three methods supporting focused outlier detection: Method 1 presented just for better explaining the methodology, Method 2 already used in [15] and new Method 3 proposed for outlier document detection in this paper. Section 5 presents the application of the proposed outlier detection method in the domain of autism, explored in our previous research [15], justifying our approach by showing the reduction of the search space of documents in the phase of outlier document detection (by proposed Method 3), and by proposing an approach to narrowing down the search space of potential bridging concepts through automated keyword-based representation of outlier documents (by proposed Method 4). Evaluation of the proposed cross-context knowledge discovery methodology on the migraine–magnesium domain pair, explored originally in Swanson's research [17], is provided in Section 6, showing that by using the proposed methodology we can easily reverse-engineer his discoveries. Section 7 presents other related work in literature mining, and Section 8 concludes by a summary and directions for further work.

2. BISOCIATIVE LINK DISCOVERY AND SWANSON'S ABC MODEL

This paper presents an approach to computational knowledge discovery through the mechanism of bisociation. Bisociative reasoning is at the heart of creative, accidental discovery (e.g. serendipity) and is focused on finding unexpected links by crossing contexts. This section establishes the correspondence between the Koestler's [12] and the Swanson's [17] creative knowledge discovery approaches.

2.1. Koestler's creativity model

In order to connect seemingly unrelated information, scientific discovery requires creative thinking, for example, by using

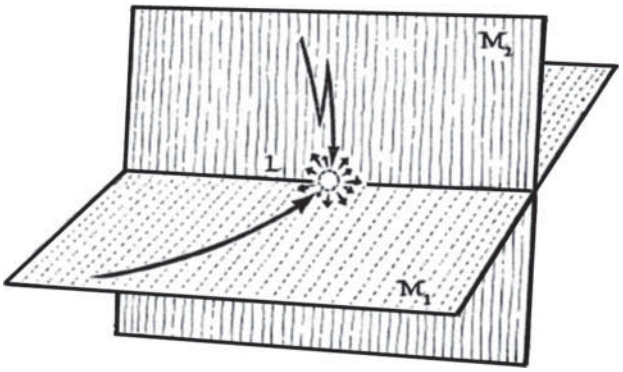


FIGURE 1. Koestler's schema of bisociative discovery [12].

metaphors or analogies between concepts from different domains. These modes of thinking allow the mixing of conceptual categories or contexts, which are normally separated. One of a functional basis for these modes is the idea of bisociation, coined by Koestler [12]:

> *'The pattern ... is the perceiving of a situation or idea, L, in two self-consistent but habitually incompatible frames of reference, M1 and M2. The event L, in which the two intersect, is made to vibrate simultaneously on two different wavelengths, as it were. While this unusual situation lasts, L is not merely linked to one associative context but bisociated with two.'*

Koestler found bisociation to be the basis for human creativity in seemingly diverse human endeavors, such as humor, science and arts. The concept of bisociation is illustrated in Fig. 1.

In this paper we explore a specific pattern of bisociation: links between concepts which belong to different contexts. The creative act is to find links which lead 'out-of-the-plane' in Koestler's terms, i.e. links which cross two or more different domains. More precisely, we claim that two concepts are bisociated if and only if:

(i) there is no direct, obvious evidence linking them, and
(ii) one has to cross contexts to find the link, and
(iii) this new link provides some novel insight into the problem domain.

Although Koestler's insight into creative knowledge discovery is rather old, no comprehensive ICT methodology has yet been developed on this basis. Our work contributes to the idea of building a new ICT methodology based on Koestler's observations, investigated within a European FP7 FET-Open project BISON (2008–2011, http://www.bisonet.eu/). Opposed to our research, which is based on outlier document detection and text mining, the main body of research in the BISON project focuses on graph mining.

2.2. Swanson's ABC model

To support creative literature-based discovery in medical domains, Swanson has designed the *ABC model* approach [17] which investigates whether an agent A is connected with a phenomenon C by discovering complementary structures via interconnecting phenomena B. Two literatures are complementary if one discusses the relations between A and B, while a disparate literature investigates the relations between B and C. If combining these relations suggests a previously unknown meaningful relation between A and C, this can be viewed as a new piece of knowledge that may contribute to a better understanding of phenomenon C.

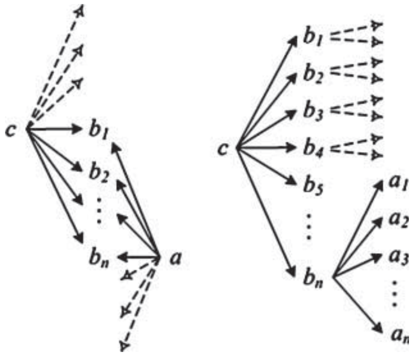


FIGURE 2. Closed (left) and open (right) discovery process as defined by Weeber *et al*. [14].

Weeber *et al*. [14] defined hypothesis generation as an *open discovery process*, and hypothesis testing as a *closed discovery process*. In an open discovery process only the phenomenon under investigation (C) is given in advance, while the target phenomenon A is still to be discovered. In a closed discovery process, both C and A are known and the goal is to search for bridging concepts in B in order to support the validation of the hypothesis about the connection between A and C.

Smalheiser and Swanson [18] developed an online system named ARROWSMITH, which takes as input two sets of titles from disjoint domains A and C and lists bridging terms (b-terms) that are common to literatures A and C; the resulting b-terms are used to generate novel scientific hypotheses. As stated by Swanson *et al*. [19], the major focus in literature-based discovery has been on the closed discovery process, where both A and C have to be specified in advance.

The difference between closed and open discovery process is illustrated in Fig. 2, where uppercase symbols A, B and C represent sets of terms (e.g. literature or collection of records), while lowercase symbols a, b and c represent single terms.

2.3. Unifying the Koestler's and Swanson's models

In this paper we explore an alternative approach to bisociative cross-context link discovery, based on outlier documents, used to detect implicit relationships across different pre-defined domains of expertise. It can be seen that—in terms of the Swanson's ABC model used in literature mining—this is an approach to closed knowledge discovery, where two domains of interest

TABLE 1. Unifying Koestler's and Swanson's models of creative knowledge discovery.

Koestler's model	Swanson's model
Bisociative link discovery process	Closed discovery process
Frames of reference (contexts) $M1$ and $M2$	Domains of interest A and C
Bisociative cross-context link L in $M1 \cap M2$	Bridging b-term in terms$(A) \cap$ terms(C)

A and C are identified by the expert prior to starting the knowledge discovery process. In terms of the Koestler's model, the two domains A and C correspond to the two habitually incompatible frames of reference $M1$ and $M2$. Moreover, linking b-terms that are common to literatures A and C, explored by Smalheiser and Swanson [18], clearly correspond to Koestler's notion of a situation or idea L, which is not merely linked to one associative context but bisociated with two contexts $M1$ and $M2$. These observations are summarized in Table 1.

3. SETTING THE STAGE FOR USING OUTLIERS IN CREATIVE LITERATURE MINING

Creative thinking constantly involves a process of evoking latent possibilities to discover new useful information and unforeseen knowledge. The fundamental reason for our focus on outliers lies in the associative creativity theory [13]: Mednick defined creative thinking as the ability to generate new combinations of distant associative elements. Additionally, we base our approach on the view that marginal observations are not necessarily characterized by mistakes or inaccuracies but that outliers [20] may provide valuable information. Besides being distributed far from the data mass, outliers are by definition rare occurrences that represent a very low fraction of total data.

The rationale for exploring the rarity principle in knowledge discovery is that if a piece of information is abundant in the set of articles, it may be speculated that its impact to the field under study is well-covered; however, if it appears rarely, not many researchers are acquainted with it, so it might be worth exploring it further. Similarly to dispersal-limited species from ecology, such pieces of information might be either on their way to extinction or might embody a potential for new developments in the field. In order to distinguish between the two options, expert guidance is needed in the process. In our approach, human involvement assures that the search process concentrates on those parts of the search space that are interesting and meaningful for a subject expert. Expert's explicit involvement in the process enables more focused and faster search for results that are worthwhile for further investigation.

Rarity of terms as means for open knowledge discovery has been explored already in the RaJoLink method [15]. The method can be used to find interesting scientific articles in the MEDLINE database [21], to compute different statistics and to analyze the articles with the aim to discover new knowledge. The RaJoLink method involves three principal steps, Ra, Jo and Link, which have been named after the key elements of each step: rare terms, joint terms and linking terms, respectively. In step Ra, interesting rare terms in literature about the phenomenon C under investigation are identified. In step Jo, all available articles about the selected rare terms are inspected and interesting joint terms that appear in the intersection of the literatures about rare terms are identified as the candidates for A. This results in a candidate hypothesis that C is connected with A. In order to provide explanation for hypotheses generated in step Jo, in the Link step the method searches for b-terms, linking the literature on joint term a (a-term) and the literature on term c (c-term). Note that, as illustrated in Fig. 2, steps Ra and Jo implement the open discovery, while step Link corresponds to the closed discovery process, searching for b-terms when A and C are both already known. The methodological description of the three steps has been provided in our previous publications [15, 22, 23].

In this paper we focus on the closed discovery process, addressed in the Link step of the RaJoLink methodology. The research presented in this paper aims at finding cross-context links between concepts from two disparate literature sources A and C, based on exploring outlier articles of the two domains. Our method assumes that by exploring outlier documents it will be easier to discover linking b-terms (bridging concepts) that establish previously unknown links between literatures A and C.

4. EXPLORING OUTLIERS IN CLOSED KNOWLEDGE DISCOVERY

Creative thinking often requires focusing on problems from new perspectives with the ability to bridge the gap between different contexts. Such relations between distinct contexts can be revealed through the bridging concepts. Since this may lead to the generation of many possible ideas, the innovative composition of hypotheses as well as support for facilitated exploration of alternatives are needed for creative knowledge discovery.

Based on this assumption, we have experimented with three methods for outlier detection used in our approach to closed knowledge discovery, where outliers are used as heuristic guidance to speed up the search for bridging concepts between different domains of expertise. The intuition behind this research is that outlier documents in the domain literature have a higher probability to provide observations that may lead to the discovery of bridging concepts. In this way the outliers can be employed for finding new interesting relations among the dispersed literatures of different domains.

Let us focus on outlier detection as the main step of the proposed closed discovery process. The closed discovery process is in this work supported by using the OntoGen tool for semi-automatic topic ontology construction [24].[1] One of its features is its facility of visualizing the similarity between the selected documents of interest. Similarity between documents can be determined by calculating the cosine of the angle between two documents represented as *Bag of Words* (*BoW*) vectors, where the Bag of Words approach [25, 26] is used for representing a collection of words from text documents disregarding grammar and word order. Content similarity is measured using the standard *TFIDF* (term frequency inverse document frequency) weighting method [27]. The standard BoW approach is used together with the TFIDF weighting. BoW representation of text documents is employed for extracting words with similar meaning. In the BoW vector space representation, each word from the document vocabulary stands for one dimension of the multidimensional space of text documents. Corpus of text documents is then visualized in form of TFIDF vectors, where each document is encoded as a feature vector with word frequencies as elements.[2]

All the documents are then sorted according to their similarity, which will enable the use of OntoGen to compare the terms in content-related documents by comparing neighboring documents from the list. 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 order the documents according to their similarity to the representative document (centroid) of a selected domain. Documents ordered based on the cosine similarity measure can be visualized in OntoGen by a similarity graph illustrated in Fig. 3. Cosine similarity values fall within the [0, 1] interval. Value 0 represents extreme dissimilarity, where two documents (a given document and the centroid vector of its cluster) share no common words, while 1 represents the similarity between two exactly identical documents in the BoW representation.

The main novelty of the proposed methodology is to visualize outlying (and their neighboring) documents in the documents' similarity graph in order to speed up the search for bisociations in the combined set of literatures A and C (AC in Fig. 4). Our argumentation is that outlying documents of two implicitly linked subjects can be used to search for relevant bridging concepts between the two subject domains. The idea of representing instances of literature A together with instances of literature C in the same similarity graph with the purpose of searching for their bisociative links is a unique aspect of our approach in comparison with the literature-based discovery investigated by other researchers.

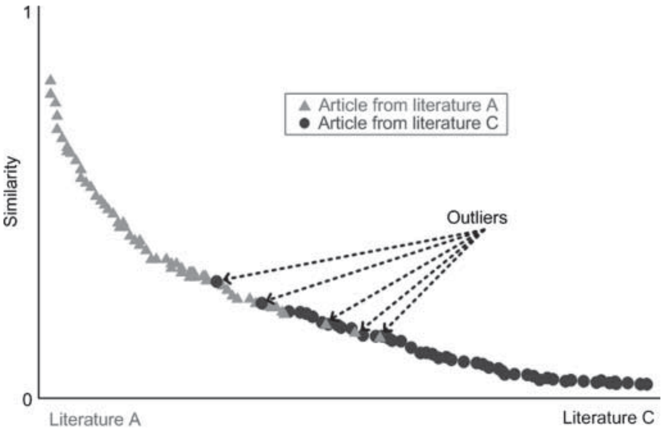


FIGURE 3. A graph representing instances of literature A and instances of literature C in terms of their content similarity. Documents are ordered according to their similarity to the representative document (centroid) of literature A, where outliers from literature C are positioned among the documents from literature A.

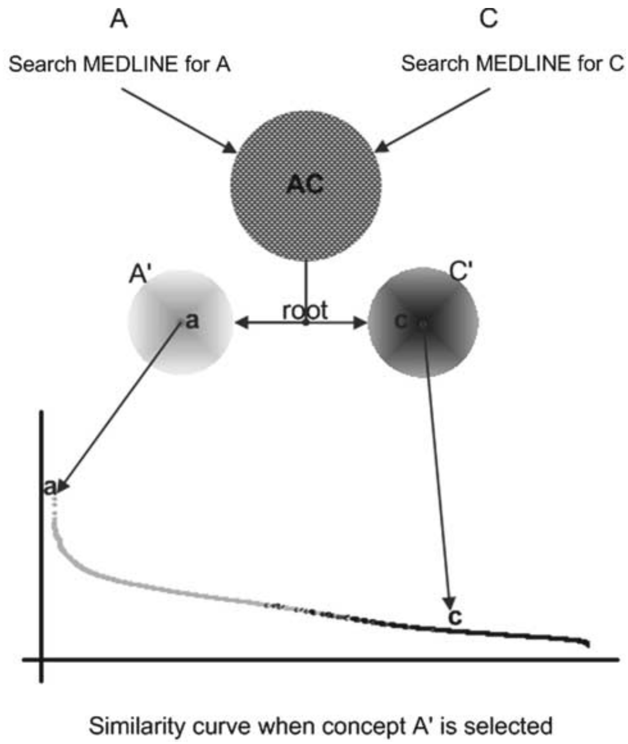


FIGURE 4. Summary of the proposed approach to cross-context link discovery when exploring two disparate literature sources A and C, using the idea of outliers. Note that AC denotes $A \cup C$, i.e. the set of documents from both domains A and C, while a and c denote centroids of domains A and C, respectively.

Having disparate literatures A and C, both domains are examined by the combined data set of literatures A and C in order to assess whether they can be connected by implicit relations. In the closed discovery process, our approach assumes

[1] OntoGen is freely available for download from http://OntoGen.ijs.si/.

[2] Elements of vectors are weighted with the TFIDF weights as follows [28]: the ith element of the vector containing frequency of the ith word is multiplied with $\mathrm{IDF}_i = \log(N/\mathrm{df}_i)$, where N represents the total number of documents and df_i is document frequency of the ith word (i.e. the number of documents from the whole corpus in which the ith word appears).

that by exploring outlier documents it will be easier to discover linking b-terms that bridge literature A and literature C. The closed discovery setting and the use of exploring outliers as means for discovering bridging concepts acting as cross-context links are summarized in Fig. 4, which is used to explain the distinction between the three methods outlined below.

> *Method* 1.[3] The first method uses domains A and C, and builds a joint document set AC (i.e. $A \cup C$). For this purpose, two individual sets of documents (e.g. titles, abstracts or full texts of scientific articles), one for each term under research (literature A and literature C), are automatically retrieved from MEDLINE or extracted from other document sources. The documents from the two individual sets are loaded in a single text file (i.e. a joint document set AC) where each line is interpreted as a document with the first word in the line being its title. We consider all the words 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. From a joint document set $A \cup C$, a similarity graph between two document sets A and C is constructed by OntoGen. In this case, in Fig. 4, $A' = A$ and $C' = C$. In this setting, centroid $a$of A' is the centroid of document set A, and centroid c of C' is the centroid of document set C, and the similarity graph is built by ranking and visualizing all the documents from AC in terms of their similarity to centroid a of document set A.
>
> *Method 2*.[4] The second method again uses domains A and C, builds a joint document set AC, but then the OntoGen tool is used to build two document clusters, A' and C' (where $A' \cup C' = AC$), using OntoGen's 2-means clustering algorithm. In this case, the similarity graph of Fig. 4 is built by ranking and visualizing all the documents from AC in terms of their similarity to centroid a of cluster A'.

Dividing the documents of the two literatures into two clusters, as suggested by Method 1 and Method 2, is a rather straightforward solution, but it requires substantial expert involvement in the process of exploring potential b-terms in many potential outlier documents.

The question whether one can reduce the number of potential outlier documents by including additional mechanisms motivated the development of Method 3, which combines the strategies employed in Methods 1 and 2.

> *Method 3*.[5] The third method starts by first employing Method 2 to construct two document clusters A' and C'. Then, based on domains A and C, each cluster is further divided into two document subclusters. This step is similar to Method 1 applied on each individual document cluster: cluster A' is divided into subclusters $A' \cap A$ and $A' \cap C$, while cluster C' is divided into $C' \cap A$ and $C' \cap C$. In this method, subclusters $A' \cap C$ (outliers of C, consisting of documents of domain C only) and $C' \cap A$ (outliers of A, consisting of documents of domain A only) are the two document sets used for forming the similarity graph of Fig. 4 used for identifying outlier documents.

Note that in Fig. 4 the original domain of interest is domain A and the similarity graph is constructed from the point of view of centroid of A' as in Method 2 (see also Fig. 5), while in the two other methods, the similarity graph is constructed from the point of view of centroid a of domain A in Method 1, and centroid of $C' \cap A$ (or$A' \cap C$) in Method 3 (see also Fig. 6). In the next two sections we present our experiments and provide additional explanations of the methods.

5. OUTLIER AND *b*-TERM DETECTION IN THE AUTISM–CALCINEURIN DOMAIN PAIR

A straightforward approach to finding outlier documents is to apply Method 1, which can be considered as a supervised clustering approach, where the two clusters correspond to domains A and C, which are the topic of our investigation.[6] In this experimental setting, when exploring the literature on *autism* (domain A), the collaborating medical expert has proposed to take *calcineurin* literature (domain C) as the second target domain. We omit the presentation of the results of Method 1, as they are very similar to the results of Method 2, which will also serve as a starting point of the best performing Method 3. The application of Method 2 on the autism–calcineurin documents pair results in Fig. 5.

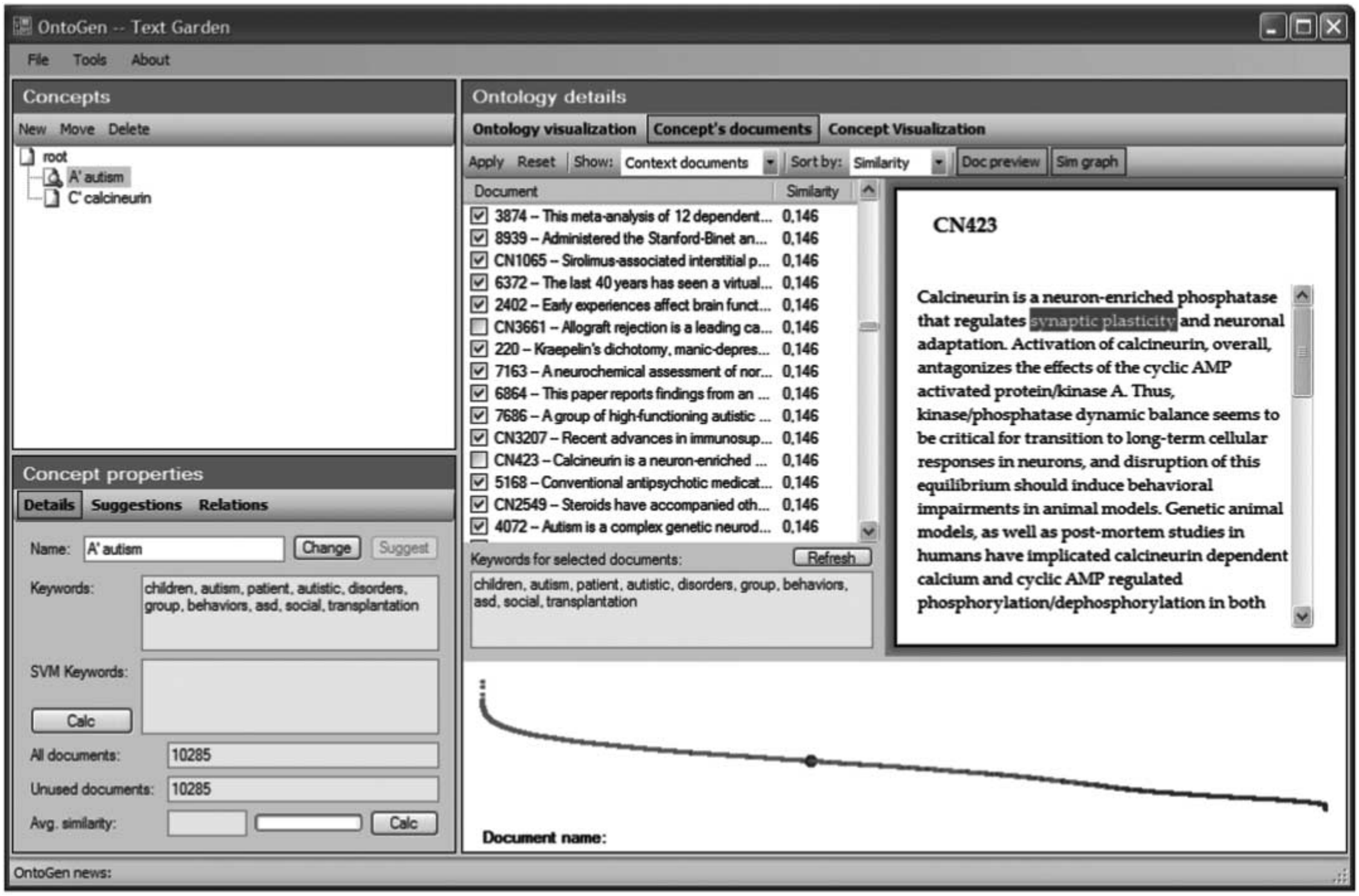
Figure 5 shows the similarity graph representing instances of literature A (*autism* cluster A') and instances of literature C (*calcineurin* cluster C') according to their content similarity, where A denotes the set of documents about autism, C denotes the set of documents about calcineurin, and A' and C' denote the two clusters formed by OntoGen's 2-means clustering (A' consisting mostly of documents from domain A, and C' consisting mostly of documents from domain C). Two main article topics (A' *autism cluster* and C' *calcineurin cluster*) are listed on the left side of the window. As the autism topic is selected, the list of abstracts of scientific articles, which are in the relationship with this selected topic, is presented in the central part of the OntoGen's window. The documents detected as outliers are visualized and their content is presented on the screen by simply clicking on the pixel representing

[3] Method 1 is introduced in this paper mainly for the sake of conceptualizing and better explaining the proposed approach.

[4] Method 2 is a reformulation of the method implemented in RaJoLink [15].

[5] Method 3 is a new method proposed in this paper.

[6] A and C are symmetric in the closed discovery setting, and can be used interchangeably. For simplicity, we have therefore used symbol A for autism, and C for calcineurin.



Name: A' autism

Keywords: children, autism, patient, autistic, disorders, group, behaviors, asd, social, transplantation

SVM Keywords:

All documents: 10285

Unused documents: 10285

Avg. similarity:

Show: Context documents

Sort by: Similarity

	Document	Similarity
☑	3874 – This meta-analysis of 12 dependent...	0,146
☑	8939 – Administered the Stanford-Binet an...	0,146
☑	CN1065 – Sirolimus-associated interstitial p...	0,146
☑	6372 – The last 40 years has seen a virtual...	0,146
☑	2402 – Early experiences affect brain funct...	0,146
□	CN3661 – Allograft rejection is a leading ca...	0,146
☑	220 – Kraepelin's dichotomy, manic-depres...	0,146
☑	7163 – A neurochemical assessment of nor...	0,146
☑	6864 – This paper reports findings from an ...	0,146
☑	7686 – A group of high-functioning autistic ...	0,146
☑	CN3207 – Recent advances in immunosup...	0,146
□	CN423 – Calcineurin is a neuron-enriched ...	0,146
☑	5168 – Conventional antipsychotic medicat...	0,146
☑	CN2549 – Steroids have accompanied oth...	0,146
☑	4072 – Autism is a complex genetic neurod...	0,146

Keywords for selected documents: children, autism, patient, autistic, disorders, group, behaviors, asd, social, transplantation

CN423

Calcineurin is a neuron-enriched phosphatase that regulates synaptic plasticity and neuronal adaptation. Activation of calcineurin, overall, antagonizes the effects of the cyclic AMP activated protein/kinase A. Thus, kinase/phosphatase dynamic balance seems to be critical for transition to long-term cellular responses in neurons, and disruption of this equilibrium should induce behavioral impairments in animal models. Genetic animal models, as well as post-mortem studies in humans have implicated calcineurin dependent calcium and cyclic AMP regulated phosphorylation/dephosphorylation in both

Document name:

FIGURE 5. OntoGen's similarity graph representing instances of literature A' (*autism* cluster constructed by OntoGen) and instances of literature C' (*calcineurin* cluster constructed by OntoGen) in terms of their content similarity. The distinctive article about the substance calcineurin (*CN423*) is detected among the autism cluster documents A', and visualized in the OntoGen graphical user interface. Note that a more detailed description of the OntoGen screen layout is given in the Appendix.

the document, as shown in Fig. 5, where the distinctive calcineurin article (CN423) is visualized among the autism cluster documents.

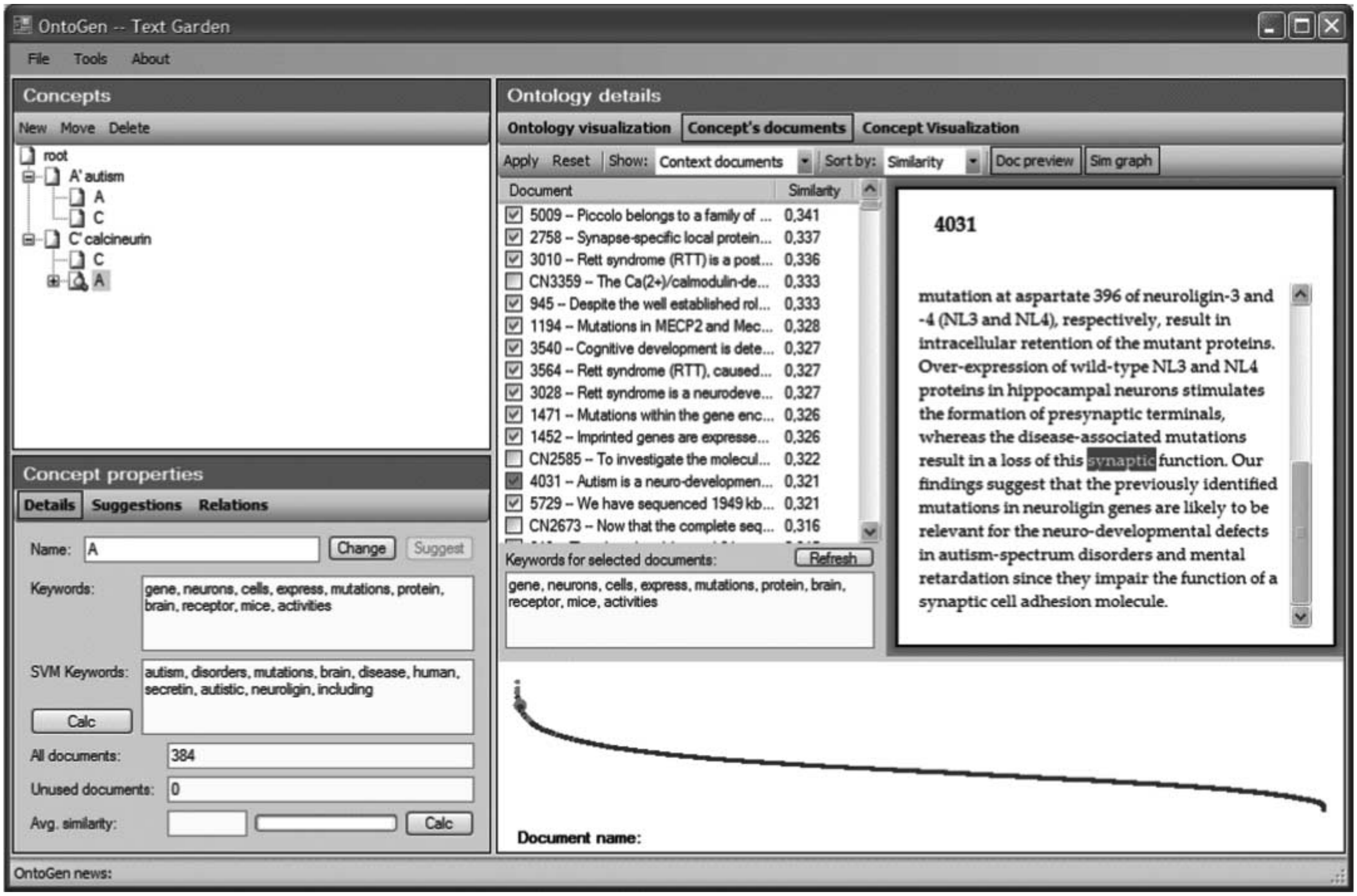
5.1. Detailed description of Method 3

At the first stage of Method 3, the OntoGen's clustering algorithm is employed to generate two clusters A' and C' (like in Method 2). Then, for each of the clusters, supervised clustering approach is applied taking into account documents' original domains A and C. In the supervised clustering approach the user has a clear idea of what the (sub-)clusters should be about. As a result, a two-level tree hierarchy of clusters that is shown in Fig. 6 (in the top left window) is generated. Our hypothesis is that by observing the similarity graphs in clusters $A' \cap C$ and $C' \cap A$ the candidates for outlier documents can be found much more effectively than by using Method 1 or Method 2. It will be shown in Section 5.2 that Method 3 indeed results in a substantial reduction of the search space of outlier documents.

5.2. Reducing the search space of potential outliers by Method 3

Theoretically, in Method 2—which ranks all the documents according to their similarity to the centroid a of cluster A'—one would need to consider as outliers all the 4958 documents of calcineurin cluster C', evaluating one by one, concentrating on the documents from C' that have in their similarity neighborhood also documents from the other cluster A' (see e.g. outlier *CN423* in the similarity graph of Fig. 5).

This search space is substantially reduced in Method 3. The effectiveness of Method 3 can be explained from the contingency table perspective, illustrated by the sizes of document sets shown in Table 2. Instead of sorting all the 15 243 documents in terms of their similarity to the centroid of A', Method 3 takes only a substantially smaller set of 4958 documents in C' to be ranked according to their similarity to the centroid of $C' \cap A$, exploring a reduced set of 384 potential outliers of A only (see 384 in $C' \cap A$ in Table 2) and concentrating on documents from $C' \cap A$ that have in



root
- A' autism
  - A
  - C
- C' calcineurin
  - C
  - A

Show: Context documents

Sort by: Similarity

	Document	Similarity
☑	5009 -- Piccolo belongs to a family of ...	0,341
☑	2758 -- Synapse-specific local protein...	0,337
☑	3010 -- Rett syndrome (RTT) is a post...	0,336
□	CN3359 -- The Ca(2+)/calmodulin-de...	0,333
☑	945 -- Despite the well established rol...	0,333
☑	1194 -- Mutations in MECP2 and Mec...	0,328
☑	3540 -- Cognitive development is dete...	0,327
☑	3564 -- Rett syndrome (RTT), caused...	0,327
☑	3028 -- Rett syndrome is a neurodeve...	0,327
☑	1471 -- Mutations within the gene enc...	0,326
☑	1452 -- Imprinted genes are expresse...	0,326
□	CN2585 -- To investigate the molecul...	0,322
☑	4031 -- Autism is a neuro-developmen...	0,321
☑	5729 -- We have sequenced 1949 kb...	0,321
□	CN2673 -- Now that the complete seq...	0,316

Keywords for selected documents: gene, neurons, cells, express, mutations, protein, brain, receptor, mice, activities

4031

mutation at aspartate 396 of neuroligin-3 and -4 (NL3 and NL4), respectively, result in intracellular retention of the mutant proteins. Over-expression of wild-type NL3 and NL4 proteins in hippocampal neurons stimulates the formation of presynaptic terminals, whereas the disease-associated mutations result in a loss of this synaptic function. Our findings suggest that the previously identified mutations in neuroligin genes are likely to be relevant for the neuro-developmental defects in autism-spectrum disorders and mental retardation since they impair the function of a synaptic cell adhesion molecule.

Name: A

Keywords: gene, neurons, cells, express, mutations, protein, brain, receptor, mice, activities

SVM Keywords: autism, disorders, mutations, brain, disease, human, secretin, autistic, neuroligin, including

All documents: 384

Unused documents: 0

Avg. similarity:

FIGURE 6. Documents from literatures A and C are clustered according to Method 3, first using unsupervised approach (autism cluster A' and calcineurin cluster C') and, on the second level, using supervised approach (A and C) within the clusters obtained in the unsupervised way. Document similarity is shown in OntoGen's graph representing instances of literature $C' \cap A$ (i.e. documents which are outliers of literature A) and instances of literature C'. The distinctive article from the autism domain (4031) is detected in the calcineurin cluster of documents C', and visualized in the OntoGen graphical user interface. The indicated document 4031 is [29].

TABLE 2. Contingency table showing the numbers of documents in reduced document sets $A' \cap C$ and $C' \cap A$.

Predicted/actual	A (autism domain)	C (calcineurin domain)	Total
A' (autism cluster)	8981	$A' \cap C$ 1304	10 285
C' (calcineurin cluster)	$C' \cap A$ 384	4574	4958
Total	9365	5878	15 243

their similarity neighborhood also documents from cluster C'. Alternatively, 1304 potential outliers of C from $A' \cap C$ are explored within 10 589 documents in A' in the similarity graph, where these documents are ranked according to their similarity to the centroid of $A' \cap C$.

Compared to the document sets that need to be explored in Method 2, the contingency table shown in Table 2 clearly shows the order of magnitude of Method 3 document sets reductions.

This set reduction is further shown by a graphical representation of the explored document sets in Fig. 7. The figure illustrates the two original classes of documents, the autism (A) and calcineurin (C) domain, the autism cluster A' and the calcineurin cluster C' generated by OntoGen using 2-means clustering, as well as document set intersections $C' \cap A$ (containing only documents of class autism) and $A' \cap C$ (containing only documents of class calcineurin).

Consider document 4031 in $C' \cap A$. This is one of autism documents: an outlier document which is—according to the OntoGen similarity measure—closer to centroid c of C' than to centroid a of A' and is therefore clustered into C'. In this

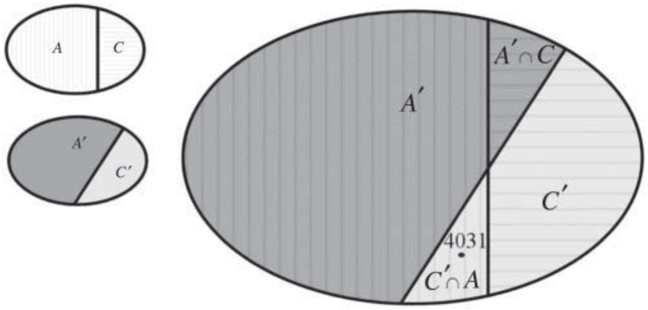


FIGURE 7. The reduced document sets explored in Method 3: $C' \cap A$ and $A' \cap C$. Sets A and C are depicted by patterns of vertical and horizontal lines, respectively, while sets A' and C' are depicted in different colors. Note that the set $C' \cap A$ represents the outliers of documents A and the set $A' \cap C$ represents the outliers of documents C.

document the *b*-term *synaptic*—linking the two literatures A and C—was found, which was confirmed by the medical expert as a relevant linking term [23].

5.3. Reducing the search space of potential *b*-terms by Method 4

By using Method 3 it was relatively easy to find *b*-term *synaptic* in document 4031 as the document is short (it is an abstract). Consider, however, a longer document (e.g. a full size document) to be explored; in this case finding a *b*-term in the outlier document can be a very time consuming effort. As will be shown in this section, Method 4—which is an extension of Method 3—can improve also the efficiency of searching for linking *b*-terms when exploring links between literatures A and C.

Note that Srinivasan *et al.* [30], who developed an algorithm for bridging concept identification that is declared to require the least amount of manual work in comparison with other studies, still needed substantial time and human effort for collecting evidence relevant to the hypothesized connections. In comparison, one of the advantages of our approach is that the domain expert needs to be involved only in exploring the potential *b*-terms in outlier documents, instead of exploring all the most frequent potential *b*-terms in all the documents. To ensure the efficiency of *b*-term discovery, the OntoGen similarity graph visualization and a list of OntoGen keywords describing an outlier document have proved to be very effective. Therefore, the necessary human effort can be substantially reduced by following the guidance from our approach described below. In the key steps of the discovery process, our methodology supports the expert by listing documents by their content similarity and by sorting term candidates by frequencies which turned out to be a useful estimate for their potential for knowledge discovery [31]. The final choice of *b*-terms is then provided by a domain expert.

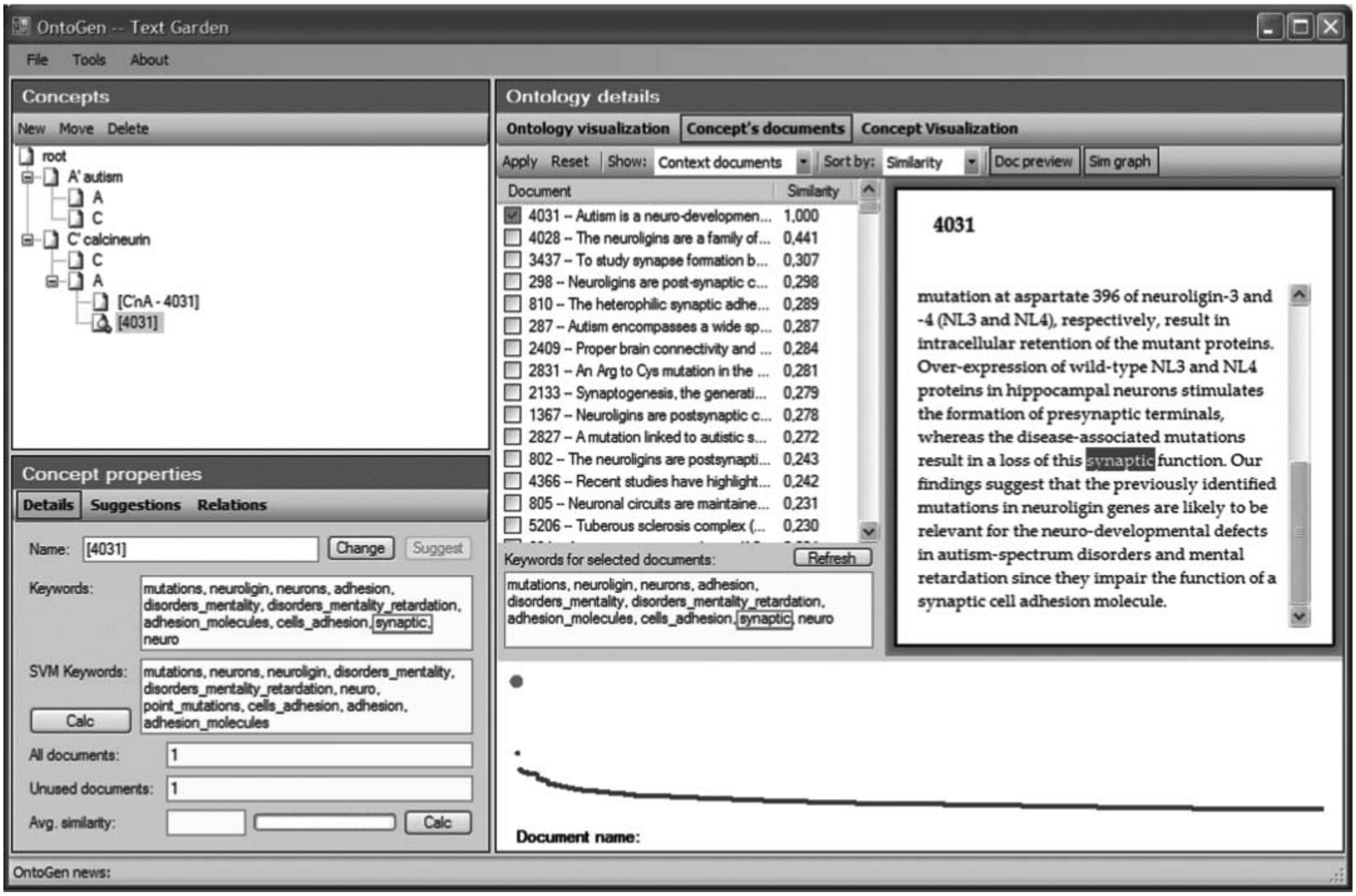
To improve the efficiency of searching for linking *b*-terms when seeking for terms that connect literatures A and C, we have again used OntoGen: this time OntoGen is not used for finding outlier documents but for automatically suggesting the potential *b*-terms which appear in the selected outlier document. To this end, consider the outlier document 4031 as a singleton subcluster {4031} of $C' \cap A$. This enables OntoGen to describe the documents keywords: *mutations, neuroligin, neurons, adhesion, disorders_mentality, disorders_mentality_retardation, adhesion_molecules, cells_adhesion, synaptic, neuro.* Note that the list of keywords includes *synaptic* as one of the potential *b*-terms connecting domains A and C. This $b-$term detection step of the closed discovery process is illustrated in Fig. 8.

This study shows that the proposed methodology, consisting of Method 3 for outlier document detection and Method 4 for *b*-term detection can be successfully used to support creative discovery of bridging concepts among different domains, which are usually not considered together. Outlier documents are able to indicate the existence of terms (i.e. bridging concepts), which can be used to form bisociations. Thus the identification of outlier documents can indeed improve the closed discovery process.

5.4. Medical interpretation of the discovery

MEDLINE, the biomedical bibliographic database with more than 18 million citations, provides no direct evidence of the role of calcineurin in the autism phenomena. While query *autism* to MEDLINE database gives over 15 000 hits and query *calcineurin* gives over 7000 hits, composed query *autism and calcineurin* gives no results. Thus we can look at autism literature (A) and calcineurin literature (C) as two different domains. Note, however, that query *autism and synaptic plasticity* results in 37 bibliographic references and *calcineurin and synaptic plasticity* in 116 bibliographic references,[7] indicating the *synaptic plasticity* is indeed a potential linking concept. Already in the evaluation of RaJoLink results presented in [23], the medical expert confirmed that indirect relations via discovered linking terms draw attention to interesting connections between two well developed, but not sufficiently connected, fields. In particular, she justified this statement by the following comment: '*Calcineurin is a calcium/calmodulin-dependent protein phosphatase* [32]. *Recent studies indicate that it participates in intracellular signaling pathways, which regulate synaptic plasticity and neuronal activities* [33]. *An impaired synaptic plasticity is thought to be also a consequence of the lack of FMR1 protein in fragile X syndrome which is one of the identified causes of autism* [34].' Although more thorough medical investigations will be needed to evaluate the importance of the revealed relation, it is certain that the result of literature mining with our method is surely capable of indicating

[7] http://www.ncbi.nlm.nih.gov/sites/entrez (accessed on 2 June 2010).



root
A' autism
A
C
C' calcineurin
C
A
[C'nA - 4031]
[4031]

Show: Context documents
Sort by: Similarity

Document	Similarity
☑ 4031 -- Autism is a neuro-developmen...	1,000
□ 4028 -- The neuroligins are a family of...	0,441
□ 3437 -- To study synapse formation b...	0,307
□ 298 -- Neuroligins are post-synaptic c...	0,298
□ 810 -- The heterophilic synaptic adhe...	0,289
□ 287 -- Autism encompasses a wide sp...	0,287
□ 2409 -- Proper brain connectivity and ...	0,284
□ 2831 -- An Arg to Cys mutation in the ...	0,281
□ 2133 -- Synaptogenesis, the generati...	0,279
□ 1367 -- Neuroligins are postsynaptic c...	0,278
□ 2827 -- A mutation linked to autistic s...	0,272
□ 802 -- The neuroligins are postsynapti...	0,243
□ 4366 -- Recent studies have highlight...	0,242
□ 805 -- Neuronal circuits are maintaine...	0,231
□ 5206 -- Tuberous sclerosis complex (...	0,230

4031

mutation at aspartate 396 of neuroligin-3 and -4 (NL3 and NL4), respectively, result in intracellular retention of the mutant proteins. Over-expression of wild-type NL3 and NL4 proteins in hippocampal neurons stimulates the formation of presynaptic terminals, whereas the disease-associated mutations result in a loss of this synaptic function. Our findings suggest that the previously identified mutations in neuroligin genes are likely to be relevant for the neuro-developmental defects in autism-spectrum disorders and mental retardation since they impair the function of a synaptic cell adhesion molecule.

Keywords for selected documents: mutations, neuroligin, neurons, adhesion, disorders_mentality, disorders_mentality_retardation, adhesion_molecules, cells_adhesion, synaptic, neuro

Name: [4031]

Keywords: mutations, neuroligin, neurons, adhesion, disorders_mentality, disorders_mentality_retardation, adhesion_molecules, cells_adhesion, synaptic, neuro

SVM Keywords: mutations, neurons, neuroligin, disorders_mentality, disorders_mentality_retardation, neuro, point_mutations, cells_adhesion, adhesion, adhesion_molecules

All documents: 1

Unused documents: 1

Avg. similarity:

FIGURE 8. Term *synaptic*, appearing in the keywords list of OntoGen, as a potential *b*-term discovered in document 4031.

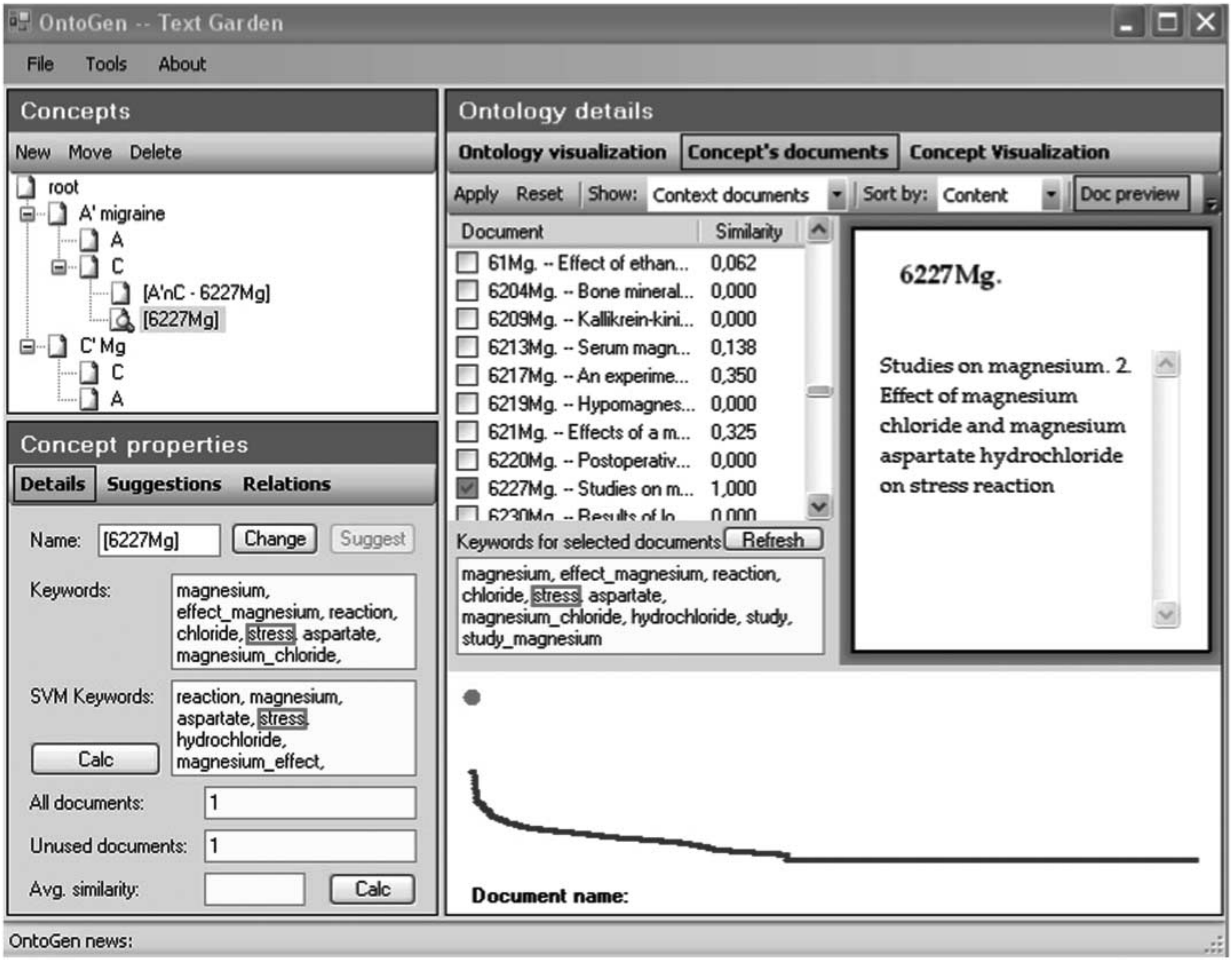
directions of further medical research recognized as relevant by the medical expert.

In Section 2.1 we claimed two concepts to be bisociated if (1) there is no direct, obvious evidence linking them, (2) one has to cross contexts to find the link and (3) this new link provides some novel insight into the problem domain. According to this claim, based on the justification in the previous paragraph, we see the autism–calcineurin bisociative relation was established via the discovered linking term, and that the presented method is effective for creative link discovery, paving a way to novel bisociations.

6. EVALUATION ON THE MIGRAINE–MAGNESIUM DOMAIN PAIR

To evaluate the proposed literature outlier detection and *b*-term extraction approach, we applied the proposed method to explore another important application domain. We investigated the ability of the proposed approach to detect the relationships between migraine and magnesium literatures. To this end, we replicated the early Swanson's migraine–magnesium experiment that represents a gold standard for the literature-based discovery. Therefore, the evaluation procedure used in this experiment differs from the original Swanson's method and the RaJoLink method in that the domain expert was not involved in the experiment.

Similar to Swanson in his original study of the migraine literature [35] we used titles as input to our closed discovery process. With the automatic support of the OntoGen tool, we performed the experiment on a subset of MEDLINE titles of articles that were published before 1988 (i.e. before Swanson's literature-based discovery of the migraine–magnesium relation) that we have retrieved with the search of the phrase: *migraine NOT magnesium*. As a result we got 6156 (migraine articles) and 6199 (magnesium articles) titles of MEDLINE articles that we analyzed by OntoGen in the closed discovery setting. Unlike Swanson we focused on those titles from migraine and magnesium literatures identified by OntoGen as interesting outliers in either the migraine or magnesium literature. Furthermore, by isolating an individual outlier document and exploring its keywords representation, we



- root
- A' migraine
  - A
  - C
    - [A'nC - 6227Mg]
    - [6227Mg]
- C' Mg
  - C
  - A

Show: Context documents
Sort by: Content

Document	Similarity
□ 61Mg. -- Effect of ethan...	0,062
□ 6204Mg. -- Bone mineral...	0,000
□ 6209Mg. -- Kallikrein-kini...	0,000
□ 6213Mg. -- Serum magn...	0,138
□ 6217Mg. -- An experime...	0,350
□ 6219Mg. -- Hypomagnes...	0,000
□ 621Mg. -- Effects of a m...	0,325
□ 6220Mg. -- Postoperativ...	0,000
☑ 6227Mg. -- Studies on m...	1,000
□ 6230Mg. -- Results of lo...	0,000

Keywords for selected documents: magnesium, effect_magnesium, reaction, chloride, stress, aspartate, magnesium_chloride, hydrochloride, study, study_magnesium

6227Mg.

Studies on magnesium. 2. Effect of magnesium chloride and magnesium aspartate hydrochloride on stress reaction

Name: [6227Mg]

Keywords: magnesium, effect_magnesium, reaction, chloride, stress, aspartate, magnesium_chloride,

SVM Keywords: reaction, magnesium, aspartate, stress, hydrochloride, magnesium_effect,

All documents: 1

Unused documents: 1

Avg. similarity:

Document name:

FIGURE 9. OntoGen's similarity graph representing instances of literature A (migraine) and instances of literature C (magnesium) in terms of their content similarity. The distinctive article about magnesium (*6227Mg*) was first detected as an outlier among the migraine documents, and then, as shown in this figure, discriminated against the rest of documents in the explored $A' \cap C$ cluster consisting of magnesium documents only. Note that b-term *stress* appears in the two OntoGen's keyword lists describing the outlier document.

gained additional information about the method's capability of detecting interesting bridging concepts (b-terms) in this large text collection.

Magnesium deficiency has been shown in several studies to cause migraine headaches (e.g. [36–40]). In the literature-based closed discovery process Swanson managed to find more than 60 pairs of articles connecting the migraine domain with the magnesium deficiency via several bridging concepts. His closer inspection of the literature about migraine and the literature about magnesium showed that 11 pairs of documents, when put together, provided confirmation of a hypothesis that magnesium deficiency may cause migraine headaches [36].

Using our methodology, we managed to identify several outlying documents in the training data sets collection (documents published before 1988) that have led us to discover important bridging concepts between the migraine and magnesium literatures e.g. *stress*, as demonstrated in Fig. 9.

Note that stress was also one of the meaningful linking terms found in the original Swanson's experiment [36]. These results prove that our methodology supports the creative discovery of bridging concepts between different domains, which are usually not considered together. Outlying documents can help us indicating the potential b-terms (i.e. bridging concepts) which can be used to form bisociations. Consequently, the identification of outliers can improve the efficiency of the closed discovery process.

7. OTHER RELATED WORK IN LITERATURE MINING

Swanson [17] and Weeber *et al.* [14] have investigated means for finding novel interesting connections between disparate research findings which can be extracted from the published literature. They have shown that the analysis of implicit associations hidden in scientific literature can guide the hypotheses formulation and lead to the discovery of new knowledge. Their approach has been briefly introduced in Section 2.2.

Several researchers have continued Swanson's line of research. An online literature-based discovery tool called BITOLA has been designed by Hristovski *et al.* [41]. It uses association rule mining techniques to find implicit relations between biomedical terms. Weeber [42] developed Literaby, the concept-based Natural Language Processing tool. The units of analysis that are essential for the approach are Unified Medical Language System Metathesaurus concepts. The open discovery approach developed by Srinivasan *et al.* [30], on the other hand, relies almost completely on the Medical Subject Headings (MeSH). Yetisgen-Yildiz and Pratt [43] proposed a literature-based discovery system called LitLinker. It mines biomedical literature by employing knowledge-based and statistical methods. All the above systems use MeSH descriptors [44] as a representation of scientific medical documents, instead of using title, abstract or full-text words. Thus, problems arise since MeSH indexers normally use only the most specific vocabulary to describe the topic discussed in a document [44] and therefore some significant terminology from the documents' content may not be covered. The Swanson's literature-based discovery approach has been extended also by Lindsay and Gordon [45], who used lexical statistics to determine relative frequencies of words and phrases. In their open discovery approach they search for words on the top of the list ranked by these statistics. However, their approach fails when applied to Swanson's first discoveries and extensive analysis has to be based on human knowledge and judgment.

The research of Ohsawa [46] indicates that tools for indicating rare events or situations play a significant role in the process of research and discovery. From this perspective, researchers have to be sensitive to curious or rare observations of phenomena in order to provide novel possible opportunities for reasoning [47] and be aware of the powerful support that data mining tools can have for choosing meaningful scenarios [46].

Outliers actually attract a lot of attention in the research world and are becoming increasingly popular in text mining applications as well. Detecting interesting outliers that rarely appear in a text collection can be viewed as searching for needles in a haystack. This popular phrase illustrates the problem with rarity and outliers since identifying useful rare objects is by itself a difficult task [47].

The rarity principle was applied in the RaJoLink literature-based open discovery process. In our earlier work [15, 22, 23], we presented the idea of extending the Swanson's *ABC model* to handle the open discovery process with rare terms from the domain literature, employing the Txt2Bow utility from the TextGarden library [48] to compute total frequencies of terms in the entire text corpus/corpora.

8. CONCLUSIONS AND FURTHER WORK

Current literature-based approaches depend strictly on simple, associative information search. Commonly, a literature-based association is computed using measures of similarity or co-occurrence. Because of their 'hard-wired' underlying criteria of co-occurrence or similarity, association-based methods often fail to discover relevant information which is not related in obvious associative ways. Especially information related across separate contexts is hard to identify with the conventional associative approach. In such cases the context-crossing connections, called bisociations, can help generate creative and innovative discoveries.

The method presented in this paper has the potential for bisociative link discovery as it allows switching between contexts by exploring linking terms in the intersections between contexts. Similar to the Swanson's closed discovery approach [17], the search for bridging concepts consists of looking for b-terms that can be found in separate sets of records: in selected literature A and another selected literature C. However, our focus is on outliers from the two sets of records and their neighboring documents which lead to substantial search reduction. We have shown that outlier documents in the OntoGen similarity graphs yield useful information in the closed discovery setting, where connections have to be found between literatures A and C. In fact, such visual analysis can present previously unseen relations, which provide new knowledge. This is an important aspect and significant contribution of our method to literature-based discovery research.

In our experiments, the proposed methodology has succeeded in identifying useful outlier documents, containing bridging concepts (b-terms) between the literature on autism and the literature on calcineurin, which proves that a combination of two previously unconnected sets of literatures in a single content similarity graph can be very effective and useful for speeding up the detection of outlier documents (in the intersections of $A' \cap C$ or $C' \cap A$) which are semantically closer to the other context. Moreover, by further exploring the outlier document and finding the keywords that best characterize the document (OntoGen keywords) or best distinguish the document (OntoGen SVM keywords), the expert's effort needed for finding linking concepts (b-terms) is substantially reduced, as it requests exploring a much smaller set of potential *b*-terms.

This work is in line with other approaches developed in the European FP7 project BISON which aims at developing efficient graph mining methods to enable effective bisociative link discovery among disparate contexts. In further work, the method of bisociative link discovery proposed in this paper will be implemented in the BISON graph mining framework and explored as means for bisociative link discovery in a number of new domains.

FUNDING

This work was partially supported by the project Knowledge Technologies (grant P2-0103) funded by the Slovene National Research Agency, and the EU project FP7-211898 BISON (Bisociation Network for Creative Information Discovery).

ACKNOWLEDGEMENT

The authors acknowledge helpful discussions with Blaž Fortuna (main developer of OntoGen), Marta Macedoni-Lukšič (medical expert on autism) and the partners of the BISON consortium.

REFERENCES

[1] Moore, D.S. and McCabe, G.P. (1999) *Introduction to the Practice of Statistics* (3rd edn). W. H. Freeman, New York.

[2] Aggarwal, C.C. and Yu, P.S. (2005) An effective and efficient algorithm for high-dimensional outlier detection. *VLDB J.*, **14**, 211–221.

[3] Lazarevic, A., Kumar, V. and Srivastava, J. (2005) Intrusion detection: a survey. In Kumar, V., Srivastava, J. and Lazarevic, A. (eds), *Massive Computing, Managing Cyber Threats*, pp. 19–80. Springer.

[4] Singhal, A. and Jajodia, S. (2006) Data warehousing and data mining techniques for intrusion detection systems. *Distrib. Parallel Databases*, **20**, 149–166.

[5] Leung, C.K.-S., Thulasiram, R.K. and Bondarenko, D.A. (2006) An efficient system for detecting outliers from financial time series. In Bell, D. and Hong, J. (eds), *Flexible and Efficient Information Handling: Proc. 23rd British National Conf. Databases* (*BNCOD 23*), Belfast, UK, July 18–20, pp. 190–198. Springer, Berlin.

[6] IPCC (2007) *Climate Change 2007: The Physical Science Basis.* Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change, pp. 1–996. IPCC.

[7] Frei, C. and Schär, C. (2001) Detection probability of trends in rare events: theory and application to heavy precipitation in the Alpine region. *J. Clim.*, **14**, 1568–1584.

[8] Ellison, A.M. and Agrawal, A.A. (2005) The statistics of rarity. *Ecology*, **86**, 1079–1080.

[9] Carney, R.S. (1997) Basing conservation policies for the deep-sea floor on current-diversity concepts: a consideration of rarity. *Biodivers. Conserv.*, **6**, 1463–1485.

[10] Boughton, D. (2001) Paradoxes in science: a new view of rarity. *Sci. Findings*, **35**, 1–6.

[11] Feldman, R. and Sanger, J. (2007) *The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data.* Cambridge University Press, Cambridge.

[12] Koestler, A. (1964) *The Act of Creation.* Macmillan, New York, 751 pp.

[13] Mednick, S.A. (1962) The associative basis of the creative process. *Psychol. Rev.*, **69**, 220–232.

[14] Weeber, M., Vos, R., Klein, H. and de Jong-van den Berg, L.T.W. (2001) Using concepts in literature-based discovery: simulating Swanson's Raynaud–fish oil and migraine–magnesium discoveries. *J. Am. Soc. Inf. Sci. Technol.*, **52**, 548–557.

[15] Petrič, I., Urbančič, T., Cestnik, B. and Macedoni-Lukšič, M. (2009) Literature mining method RaJoLink for uncovering relations between biomedical concepts. *J. Biomed. Inform.*, **42**, 219–227.

[16] American Psychiatric Association (2000) *Diagnostic and Statistical Manual of Mental Disorders* (4th edn, text revision). American Psychiatric Association, Washington, DC.

[17] Swanson, D.R. (1986) Undiscovered public knowledge. *Libr. Q.*, **56**, 103–118.

[18] Smalheiser, N.R. and Swanson, D.R. (1998) Using ARROWSMITH: a computer-assisted approach to formulating and assessing scientific hypotheses. *Comput. Methods Programs Biomed.*, **57**, 149–153.

[19] Swanson, D.R., Smalheiser, N.R. and Torvik, V.I. (2006) Ranking indirect connections in literature-based discovery: the role of Medical Subject Headings (MeSH). *J. Am. Soc. Inf. Sci. Technol.*, **57**, 1427–1439.

[20] Barnett, V. and Lewis, T. (1994) *Outliers in Statistical Data*. Wiley, New York.

[21] MEDLINE (2009) *Fact sheet.* http://www.nlm.nih.gov/pubs/factsheets/medline.html (online), accessed on 18 December 2009.

[22] Petrič, I., Urbančič, T. and Cestnik, B. (2007) Discovering hidden knowledge from biomedical literature. *Informatica*, **31**, 15–20.

[23] Urbančič, T., Petrič, I., Cestnik, B. and Macedoni-Lukšič, M. (2007) Literature mining: towards better understanding of autism. In Bellazzi, R., Abu-Hanna, A. and Hunter, J. (eds), *Proc. 11th Conf. Artificial Intelligence in Medicine in Europe* (*AIME 2007*), Amsterdam, The Netherlands, July 7–11, pp. 217–226. Springer, Berlin.

[24] Fortuna, B., Grobelnik, M. and Mladenić, D. (2006) Semi-automatic data-driven ontology construction system. In Bohanec, M. *et al.* (eds), *Proc. 9th Int. Multi-Conf. Information Society* (*IS 2006*), Ljubljana, Slovenia, October 9–14, pp. 223–226.

[25] Sebastiani, F. (2002) Machine learning in automated text categorization. *ACM Comput. Surv.*, **34**, 1–47.

[26] Grobelnik, M. and Mladenić, D. (2005) Automated knowledge discovery in advanced knowledge management. *J. Knowl. Manag.*, **9**, 132–149.

[27] Salton, G. and Buckley, C. (1988) Term weighting approaches in automatic text retrieval. *Inf. Process. Manag.*, **24**, 513–523.

[28] Fortuna, B., Grobelnik, M. and Mladenić, D. (2005) Visualization of text document corpus. *Informatica*, **29**, 497–502.

[29] Chih, B., Afridi, S.K., Clark, L. and Scheiffele, P. (2004) Disorder-associated mutations lead to functional inactivation of neuroligins. *Hum. Mol. Genet.*, **13**, 1471–1477.

[30] Srinivasan, P., Libbus, B. and Sehgal, A.K. (2004) Mining MEDLINE: postulating a beneficial role for curcumin longa in retinal diseases. In Hirschman, L. and Pustejovsky, J. (eds), *BioLINK 2004: Linking Biological Literature, Ontologies, and Databases*, Boston, MA, May 6, pp. 33–40. Association for Computational Linguistics, Morristown, NJ.

[31] Urbančič, T., Petrič, I. and Cestnik, B. (2009) RaJoLink: a method for finding seeds of future discoveries in nowadays literature. In Rauch, J. (ed.), *Proc. 18th Int. Symp. Foundations of Intelligent Systems* (*ISMIS 2009*), Prague, Czech Republic, September 14–17, pp. 129–138, Springer, Berlin, Heidelberg.

[32] Rusnak, F. and Mertz, P. (2000) Calcineurin: form and function. *Physiol. Rev.*, **80**, 1483–1521.

[33] Qiu, S., Korwek, K.M. and Weeber, E.J. (2006) A fresh look at an ancient receptor family: Emerging roles for density lipoprotein receptors in synaptic plasticity and memory formation. *Neurobiol. Learn. Mem.*, **85**, 16–29.

[34] Irwin, S., Galvez, R., Weiler, I.J., Beckel-Mitchener, A. and Greenough, W. (2002) Brain structure and the functions of FMRI protein. In Hagerman, R.J. and Hagerman, P.J. (eds), *Fragile X Syndrome.* pp. 191–205. The John Hopkins University Press, Baltimore.

[35] Swanson, D.R.(1988) Migraine and magnesium: eleven neglected connections. *Perspect. Biol. Med.*, **31**, 526–557

[36] Swanson, D.R. (1990) Medical literature as a potential source of new knowledge. *Bull. Med. Libr. Assoc.*, **78**, 29–37.

[37] Thomas, J., Thomas, E. and Tomb, E. (1992) Serum and erythrocyte magnesium concentrations and migraine. *Magnes. Res.*, **5**, 127–130.

[38] Thomas, J., Millot, J.M., Sebille, S., Delabroise, A.M., Thomas, E., Manfait, M. and Arnaud, M.J. (2000) Free and total magnesium in lymphocytes of migraine patients—effect of magnesium-rich mineral water intake. *Clin. Chim. Acta*, **295**, 63–75.

[39] Demirkaya, S., Vural, O., Dora, B. and Topçuoğlu, M.A. (2001) Efficacy of intravenous magnesium sulfate in the treatment of acute migraine attacks. *Headache*, **41**, 171–177.

[40] Trauninger, A., Pfund, Z., Koszegi, T. and Czopf, J. (2002) Oral magnesium load test in patients with migraine. *Headache*, **42**, 114–119.

[41] Hristovski, D., Peterlin, B., Mitchell, J.A. and Humphrey, S.M. (2005) Using literature-based discovery to identify disease candidate genes. *Int. J. Med. Inform.*, **74**, 289–298.

[42] Weeber, M. (2007) Drug discovery as an example of literature-based discovery. In Dzeroski, S. and Todorovski, L. (eds), *Computational Discovery of Scientific Knowledge*. Springer, Berlin, pp. 290–306.

[43] Yetisgen-Yildiz, M. and Pratt, W. (2006) Using statistical and knowledge-based approaches for literature-based discovery. *J. Biomed. Inform.*, **39**, 600–611.

[44] Nelson, S.J., Johnston, D. and Humphreys, B.L. (2001) Relationships in medical subject headings. In Bean, C.A. and Green, R. (eds), *Relationships in the Organization of Knowledge*, pp. 171–184. Kluwer Academic Publishers, New York.

[45] Lindsay, R.K. and Gordon, M.D. (1999) Literature-based discovery by lexical statistics. *J. Am. Soc. Inf. Sci.*, **50**, 574–587.

[46] Ohsawa, Y. (2006) Chance discovery: the current states of art. *Stud. Comput. Intell.*, **30**, 3–20.

[47] Magnani, L. (2005) Chance discovery and the disembodiment of mind. In Khosla, R., Howlett, R.J. and Jain, L.C. (eds), *Proc. 9th Int. Conf. Knowledge-Based Intelligent Information and Engineering Systems* (*KES 2005*), Melbourne, Australia, September 14–16, pp. 547–553. Springer, Berlin.

[48] Grobelnik, M. and Mladenić, D. (2004) Extracting Human Expertise from Existing Ontologies. EU-IST Project IST-2003-506826 SEKT, SEKT Consortium, 17 pp.

APPENDIX

The methodology described in this article is illustrated by using the OntoGen tool [24, 28]. This appendix describes the parts of the OntoGen's user interface through the OntoGen's generic panel composition shown in Fig. A1.

Panel A shows the hierarchical structure of the constructed topic ontology. One topic represents a set of documents (a cluster of documents). The top element in the hierarchy contains all the documents and is labeled 'root'.

Panel B is used to display detailed information about the selected topic (cluster) from panel A. Also, when constructing a topic ontology, this panel is used to list and select suggested subtopics (subclusters).

In panel C all the documents belonging to the parent topic (cluster) of the selected topic (cluster) from panel A are displayed. The documents of the selected topic are labeled with the checkmark. Also, the similarity of documents to the centroid of the selected topic (cluster) is shown.

Panel D is used to display the content of the selected document from panel C.

In panel E keywords that describe the selected topic from panel A are displayed. These keywords best describe the selected topic and can be used as a guideline for an expert to describe the concept related to the selected topic.

Panel F depicts the document similarity graph. The similarity of documents is computed according to the centroid of the selected topic (cluster). In the graph, the documents from the parent topic of the selected topic from panel A are shown. This similarity graph can be used for detecting outlier documents by observing points in the graph that have different color than the majority of their neighbors.

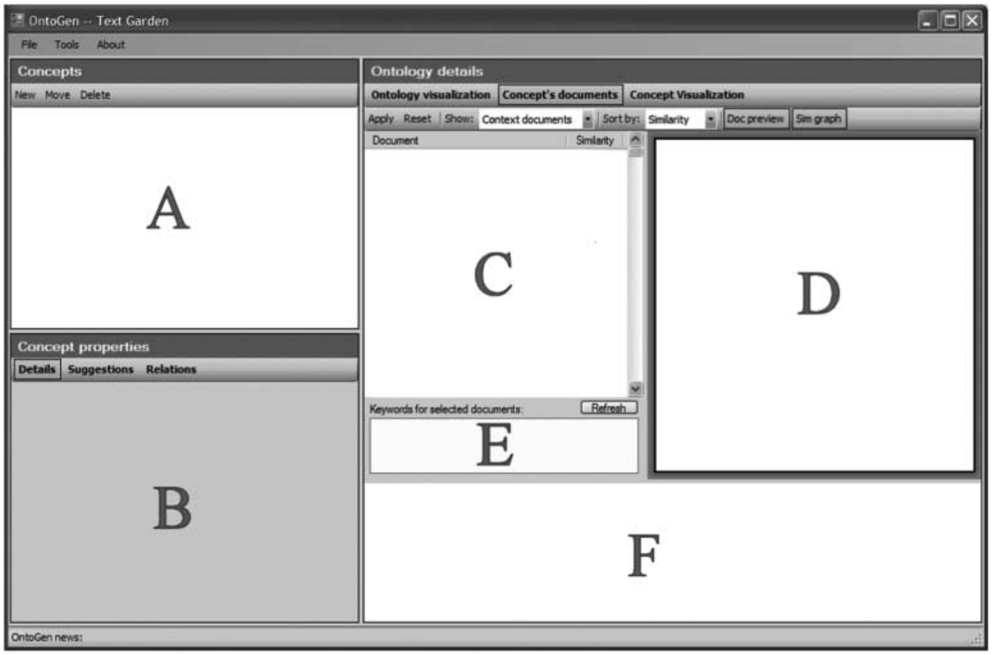


FIGURE A1. Panels in the OntoGen's screen layout.