HCI Empowered Literature Mining for Cross-Domain Knowledge Discovery

Matjaž Juršič[1, 2], Bojan Cestnik[1, 3], Tanja Urbančič[4, 1], and Nada Lavrač[1, 4]

[1] Jožef Stefan Institute, Ljubljana, Slovenia
[2] International Postgraduate School Jožef Stefan, Ljubljana, Slovenia
[3] Temida d.o.o., Ljubljana, Slovenia
[4] University of Nova Gorica, Nova Gorica, Slovenia
{matjaz.jursic, bojan.cestnik, tanja.urbancic, nada.lavrac}@ijs.si

Abstract. This paper presents an exploration engine for text mining and cross-context link discovery, implemented as a web application with a user-friendly interface. The system supports experts in advanced document exploration by facilitating document retrieval, analysis and visualization. It enables document retrieval from public databases like PubMed, as well as by querying the web, followed by document cleaning and filtering through several filtering criteria. Document analysis includes document presentation in terms of statistical and similarity-based properties and topic ontology construction through document clustering, while the distinguishing feature of the presented system is its powerful cross-context and cross-domain document exploration facility through bridging term discovery aimed at finding potential cross-domain linking terms. Term ranking based on the developed ensemble heuristic enables the expert to focus on cross-context terms with greater potential for cross-context link discovery. Additionally, the system supports the expert in finding relevant documents and terms by providing customizable document visualization, a color-based domain separation scheme and highlighted top-ranked bisociative terms.

Keywords. literature mining, knowledge discovery, cross-context linking terms, creativity support tools, human-computer interaction.

1 Introduction

Understanding complex phenomena and solving difficult problems often requires knowledge from different domains to be combined and cross-domain associations to be taken into account. These kinds of context-crossing associations are called bisociations [1] and are often needed for creative, innovative discoveries. Typically, this is a challenging task due to a trend of over-specialization in research and development, resulting in islands of deep, but relatively isolated knowledge. Scientific literature all too often remains closed and cited only in professional sub-communities. In addition, the information that is related across different contexts is difficult to identify with the associative approach, like the standard association rule learning approach [2] known from data mining and machine learning literature. Therefore, the ability of literature

mining methods and software tools for supporting the experts in their knowledge discovery process, especially in searching for yet unexplored connections between different domains, is becoming increasingly important.

The task of cross-domain literature mining has already been addressed by Swanson [3], [4], proving that bibliographic databases such as MEDLINE could serve as a rich source of hidden relations between concepts. By studying two separate literatures, the literature on migraine headache and the articles on magnesium, he discovered "Eleven neglected connections", identifying eleven linking concepts [3]. Laboratory and clinical investigations started after the publication of the Swanson's convincing evidence and have confirmed that magnesium deficiency can cause migraine headaches. This well-known example has become a golden standard in the literature mining field and has been used as a benchmark in several studies, including [5-8].

Literature mining supported discovery was successfully applied to problems, such as associations between genes and diseases [9], diseases and chemicals [10], and others. Smalheiser and Swanson [11] developed a web platform designed to assist the user in literature based discovery, which is in terms of detecting interesting cross-domain terms similar to our system. Holzinger et al. [12] describe several quality-oriented web-based tools for the analysis of biomedical literature, which include the analysis of terms (biomedical entities such as disease, drugs, genes, proteins and organs) and provide concepts associated with the given term.

Cross-domain literature mining is closely related to bisociative knowledge discovery as defined by Dubitzky et al. [13]. Assuming two domains of interest, a crucial step in cross-domain knowledge discovery is the identification of interesting bridging terms (B-terms), appearing in both literatures, which carry the potential of revealing the links connecting the two domains.

In this paper we present an online system CrossBee which helps the experts when searching for hidden links that connect two seemingly unrelated domains. As such, it supports creative discovery of cross-domain hypotheses, and could be viewed as a creativity support tool (CST). While CrossBee has been previously described [14], [15], these papers have not focused on its visual interface empowering the users in the bridging term discovery process, but have focused on its methodology and the heuristics included in the ensemble-based term ranking according to terms' bisociation potential, indicating the potential to act as bridging terms among two selected domains.

Creativity support tools are closely related to the field of human-computer interaction (HCI), as stated by Resnick et al. [16] when summarizing the aims of designing CSTs: "Our goal is to develop improved software and user interfaces that empower users to be not only more productive, but more innovative." Schneiderman [17], [18] provides a structured set of design principles for CSTs, which we follow in our implementation and use them for evaluation:

- *Support exploration.* To be successful at discovery and innovation, users should have access to improved search services providing rich mechanism for organizing search results by ranking, clustering, and partitioning with ample tools for annotation, tagging, and marking.
- *Enable collaboration.* While the actual discovery moments in innovation can be very personal, the processes that lead to them are often highly collaborative.

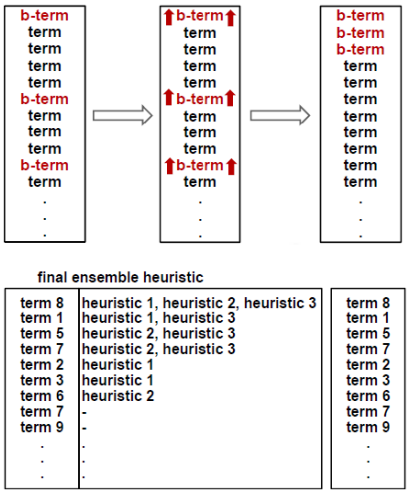
– *Provide rich history-keeping.* The benefits of rich history-keeping are that users have a record of which alternatives they have tried, they can compare the many alternatives, and they can go back to earlier alternatives to make modifications.
– *Design with low thresholds, high ceilings, and wide walls.* CST should have steep learning curve for novices (low threshold), yet provide sophisticated functionality that experts need (high ceilings), and also deliver a wide range of supplementary services to choose from (wide walls).

The main novelty of the presented system is ensemble-based ranking of terms according to their bisociative potential of contributing to novel cross-domain discoveries. This facility, together with numerous other content analysis and visualization options, distinguishes it as a powerful, user-friendly text analysis tool for cross-domain knowledge discovery support.

In the next section we present the main system functionality and a brief overview of the methodology, implemented in a contemporary workflow execution environment ClowdFlows. Section 3 presents a typical usage scenario and continues with some other system functionalities important for efficient human computer interaction in cross-context link discovery. In Section 4 we describe visual document clustering as implemented in our system. In Section 5 we summarize the most important features of the presented system and suggest some further work directions.

2 Main System Functionality and Methodology Overview

In cross-domain knowledge discovery, estimating which of the terms have a high potential for interesting discoveries is a challenging research question. It is especially important for cross-context scientific discovery such as understanding complex medi-



B-Term Identify (Analysis)

List start position: 0 Search terms: GO

You have provided 43 bterms. Out of them 43 are found in the documents. If you want to change

Pos.	Term	Votes	Inner Class Score	Documents		Hevristics' Votes				
				MIG	MAG	fr	td	rf	sv	cs	os
1	medication	6	0.9848	5	3	X	X	X	X	X	X
2	therapy	6	0.9816	106	109	X	X	X	X	X	X
3	paroxysmal [1]	6	0.9792	11	3	X	X	X	X	X	X
4	retinal	6	0.9774	10	6	X	X	X	X	X	X
5	double blind placebo	6	0.9769	5	3	X	X	X	X	X	X
6	eclampsia	6	0.9767	2	16	X	X	X	X	X	X
7	pathophysiology	6	0.9759	9	7	X	X	X	X	X	X
8	reflex	6	0.9759	8	3	X	X	X	X	X	X
9	aspect	6	0.9758	39	28	X	X	X	X	X	X
10	valve prolapse	6	0.9752	6	2	X	X	X	X	X	X
11	csf	6	0.9743	9	7	X	X	X	X	X	X
12	anti	6	0.9740	17	13	X	X	X	X	X	X
13	mitral	6	0.9736	7	3	X	X	X	X	X	X
14	blind placebo	6	0.9734	5	3	X	X	X	X	X	X
15	dysfunction	6	0.9675	6	2	X	X	X	X	X	X
16	part	6	0.9668	5	10	X	X	X	X	X	X
17	preliminary study	6	0.9635	4	6	X	X	X	X	X	X
18	irritable	5	0.9804	3	2	X		X	X	X	X
19	clinical	5	0.9787	115	90	X	X		X	X	X
20	indication	5	0.9774	3	2	X		X	X	X	X

Fig. 1. Term ranking approach (illustrated at the left) and the actual CrossBee ensemble heuristic ranking page (at the right) indicating by a cross (X) which elementary heuristics have identified the term as potential B-term.

cal phenomena or finding new drugs for yet not fully understood illnesses. Given this motivation, the main functionality of CrossBee is bridging term (B-term) discovery, implemented through ensemble-based term ranking, where an ensemble heuristic composed of six elementary heuristics was constructed for term evaluation. The ensemble-based ranking methodology, presented in more detail by Juršič et al. [14], [15], is illustrated in Fig. 1, showing the methodology of term ranking on the left and the ensemble ranked term list on the right side of the figure. The presented ranked list is the actual output produced by our system using the gold standard dataset in literature mining—the migraine-magnesium dataset [3].

We use workflow diagrams to present the cross-domain knowledge discovery methodology implemented in CrossBee. While the presented workflow diagrams are here used only as means to describe a conceptual pipeline of natural language processing (NLP) modules, the pipeline actually represent an executable workflow, implemented in the online cloud based workflow composition environment ClowdFlows [19].

The top-most level overview of the methodology, shown in Fig. 2, consists of the following steps: document acquisition, document preprocessing, outlier document detection, heuristics specification, candidate B-term extraction, heuristic terms scores calculation, and visualization and exploration. An additional ingredient shown in Fig. 2—methodology evaluation—is not directly part of the methodology, however it is an important step of the methodology development. A procedural explanation of the workflow from Fig. 2 is presented below.

1. *Document acquisition* is the first step of the methodology. Its goal is to acquire documents of the two domains, label them with domain labels and pack both domains together into the annotated document corpus (ADC) format.
2. The *document preprocessing* step is responsible for applying standard text preprocessing to the document corpus. The main parts are tokenization, stopwords labeling and token stemming or lemmatization.
3. The *outlier document detection* step is used for detecting outlier documents that are needed by subsequent heuristic specification. The output is a list (or multiple lists in case when many outlier detection methods are used) of outlier documents.
4. The *heuristic specification* step serves as a highly detailed specification of the heuristics to be used for B-term ranking. The user specifies one or more heuristics, which are later applied to evaluate the B-term candidates. Furthermore, each individual heuristic can be hierarchically composed of other heuristics; therefore an arbitrary complex list of heuristics can be composed in this step.

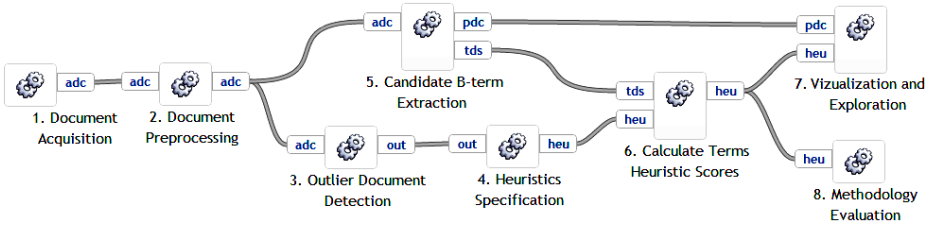


Fig. 2. Methodological steps of the cross-domain literature mining process.

5. The *candidate B-term extraction* step takes care of extracting the terms which are later scored by the specified heuristics. There are various parameters which control which kind of terms are selected from the documents (e.g., the maximal number of tokens to be joined together as a term, minimal term corpus frequency, and similar). The first output is a list of all candidate B-terms (term data set TDS) along with their vector representations. The second output is a parsed document corpus (PDC) which includes information about the input documents from ADC as well as the exact data how each document was parsed. This data is needed by the CrossBee web application when displaying the documents since it needs to be able to exactly locate specific words inside a document, e.g., to color or emphasize them.
6. *Heuristic calculation* is methodologically the most important step. It takes the list of extracted B-term candidates and the list of specified heuristics and calculates a heuristic score for each candidate term for each heuristic. The output is structurally still a list of heuristics, however now each of them contains a bisociation score for each candidate B-term.
7. *Visualization and exploration* is the final step of the methodology. It has three main functionalities. It can either take the heuristically scored terms, rank the terms, and output the terms in the form of a table, or it can take the heuristically scored terms along with the parsed document corpus and send both to the CrossBee web application for advanced visualization and exploration. Besides improved bridging concept identification and ranking, CrossBee also provides various con-

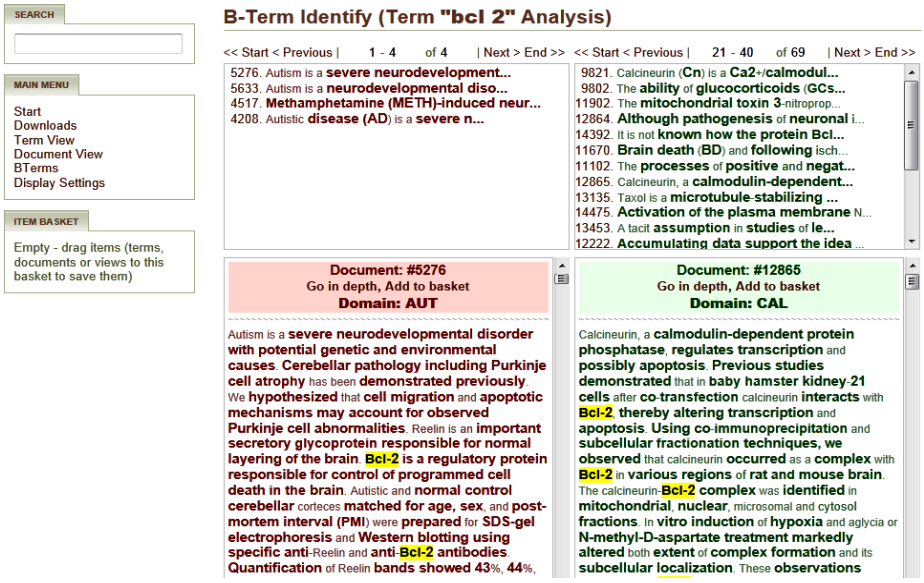


Fig. 3. One of the features most appreciated by the users is the side-by-side view of documents from the two domains under investigation. The analysis of the bcl-2 term from the autism-calcineurin domain is shown. The presented view enables efficient comparison of two documents, the left one from the autism domain and the right one from the calcineurin domain. The displayed documents were reported by Urbančič et al. [20] as relevant for exploring the relationship between autism and calcineurin.

tent presentations which further speed up the process of bisociation exploration. These presentations include side-by-side document inspection (see Fig. 3), emphasizing of interesting text fragments, and uncovering similar documents. Finally, document clustering can be used for domain exploration (see TopicCircle visualization in Fig. 4).

8. An additional *methodology evaluation* step was introduced during the development of the methodology. Its purpose is to calculate and visualize various metrics that were used to assess the quality of the methodology. Requirement to use these facilities is to have the actual B-terms as golden standard B-terms available for the domains under investigation. The methodology was actually evaluated on two problems: the standard migraine-magnesium problem well-known in literature mining, and a more recent autism-calcineurin literature mining problem.

The CrossBee system has already been successfully applied to complex domains and resulted in finding interesting cross-domain links, when replicating the results of cross-domain migraine-magnesium literature mining by Swanson [3] and replicating the results in the area of autism by Urbančič et al. [20] and Petrič et al. [21].

We are mostly interested in the CrossBee heuristics quality from the end user's perspective. Such evaluation should enable the user to estimate how many B-terms can be found among the first 5, 20, 100, 500 and 2000 terms on the ranked list of terms produced by a heuristic [14]. The ensemble heuristic, performing ensemble voting of six elementary heuristic[1], resulted in very favorable results in the training domain (migraine-magnesium domain pair), where one B-term among the first 5 terms, one B-term—no additional B-terms—among the first 20 terms, 6 B-terms—5 additional—among the first 100 terms, 22 B-terms—16 additional—among first 500 terms and all the 43 B-terms—21 additional—among the first 2000 terms. Thus, e.g., if the expert limits himself to inspect only the first 100 terms, he will find 6 B-terms in the ensemble ranked term list. These results confirm that the ensemble is among the best performing heuristics also from the user's perspective. Even though a strict comparison depends also on the threshold of how many terms an expert is willing to inspect, the ensemble is always among the best.

In the autism-calcineurin domain pair, the ensemble finds one B-term among 20 ranked terms, 2 among 100 and 3 among 500 ranked terms. At a first sight, this may seem a bad performance, but, note that there are 78,805 candidate terms which the heuristics have to rank. The evidence of the quality of the ensemble can be understood if we compare it to a simple baseline heuristic, which represents the performance achievable using random sorting of terms which appear in both domains. The baseline heuristic discovers in average only approximately 0.33 B-terms before position 2000 in the ranked list while the ensemble discovers 5; not to mention the shorter term lists where the ensemble is relatively even better compared to the baseline heuristic.

[1] The voting mechanism and the exact description of the heuristics are out of the scope of this paper; more information on the baseline, elementary and ensemble heuristics is provided by Juršič et al. [14].

The above methodology evaluation provides evidence that the users empowered with the CrossBee functionality of term ranking and visualization are able to perform the crucial actions in cross-domain discovery faster than with conventional text mining tools.

3 Typical Use Case Scenario

This section presents a typical usage scenario, illustrated with an example from the autism domain, where the aim was to find new links with calcineurin, shown in Fig. 3.

The user starts a new session by selecting two sets of documents of interest and by regulating the parameters of the system. The required input is either a PubMed query or a file with documents from the two domains, where each line contains a document with exactly three tab-separated entries: (a) document identifier, (b) domain acronym, and (c) the document text. The user can also specify the exact preprocessing options, the elementary heuristics to be used in the ensemble, outlier documents identified by external outlier detection software, the already known bisociative terms (B-terms), and others. Next, the system starts actual text preprocessing, computing the elementary heuristics, the ensemble bisociation scores and term ranking. When presented with a ranked list of B-term candidates, the user browses through the list and chooses the term(s) he believes to be promising B-terms, i.e. terms for finding meaningful connections between the two domains. At this point, the user can inspect the actual appearances of the selected term in both domains, using the efficient side-by-side document inspection.

Other functionalities of our system support the expert in advanced document exploration supporting document retrieval, analysis and visualization. The system enables document retrieval from public databases like PubMed, as well as by querying the web, followed by document cleaning and filtering through several filtering criteria. Document analysis includes document presentation in terms of statistical and similarity-based properties, topic ontology construction through document clustering, and document visualization along with user interface customization which additionally supports the expert in finding relevant documents and terms of a color-based domains separation scheme and high-lighted top-ranked bisociative terms.

A rich set of functionalities and content presentations turn our system into a user-friendly tool which enables the user not only to spot but also to efficiently investigate cross-domain links pointed out by our ensemble-based ranking methodology. Document focused exploration empowers the user to filter and order the documents by various criteria. Detailed document view provides a more detailed presentation of a single document including various term statistics. Methodology performance analysis supports the evaluation of the methodology by providing various data which can be used to measure the quality of the results, e.g., data for plotting the ROC curves. High-ranked term emphasis marks the terms according to their bisociation score calculated by the ensemble heuristic. When using this feature all high-ranked terms are emphasized throughout the whole application thus making them easier to spot (see different font sizes in Fig. 3). B-term emphasis marks the terms defined as B-terms by

the user (yellow terms in Fig. 3). Domain separation is a simple but effective option which colors all the documents from the same domain with the same color, making an obvious distinction between the documents from the two domains (different colors in Fig. 3). User interface customization enables the user to decrease or increase the intensity of the following features: high-ranked term emphasis, B-term emphasis and domain separation.

Note that the modular design of the system enabling new functionalities, in addition to the above described CrossBee functionalities; add to the fulfillment of the wide wall criterion, discussed when describing the TopicCircle document exploration facility.

4 Visual Document Clustering

Our system has the facility of clustering documents according to their similarity. 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 [22] is used for representing a collection of words from text documents disregarding grammar and word order. The BoW approach is used together with the standard *Tf·Idf* (term frequency inverse document frequency) weighting method. 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 Tf·Idf vectors, where each document is encoded as a feature vector with word frequencies as elements.[2]

The *cosine similarity*[3] measure, commonly used in information retrieval and text mining to determine the semantic closeness of two documents represented in the BoW vector space model, is used to cluster the documents. 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. For clustering, the standard k-means clustering algorithm is used.

The result of interactive top-down document clustering of the migraine-magnesium documents are presented on the left hand side of Fig. 4. At the first level, all the documents are split into one of the two domains: migraine and magnesium (top screenshot on right hand side of Fig. 4). At level 2, guided by the user, each of the two domains is further split into k sub-clusters, according to the user-selected k parameter.

[2] Elements of vectors are weighted with the Tf·Idf weights as follows: The i-th element of the vector containing frequency of the i-th word is multiplied with $Idf_i = \log(N/df_i)$, where N represents the total number of documents and df_i is document frequency of the i-th word (i.e., the number of documents from the whole corpus in which the i-th word appears).

[3] The cosine similarity is the dot product of BoW vectors, normalized by the length of the vectors: $\mathrm{CosSim}(vec_x, vec_y) = \mathrm{DotProduct}(vec_x, vec_y)/|vec_x| \cdot |vec_y|$. In the typical case, when the vectors are already normalized, the cosine similarity is identical to the dot product.

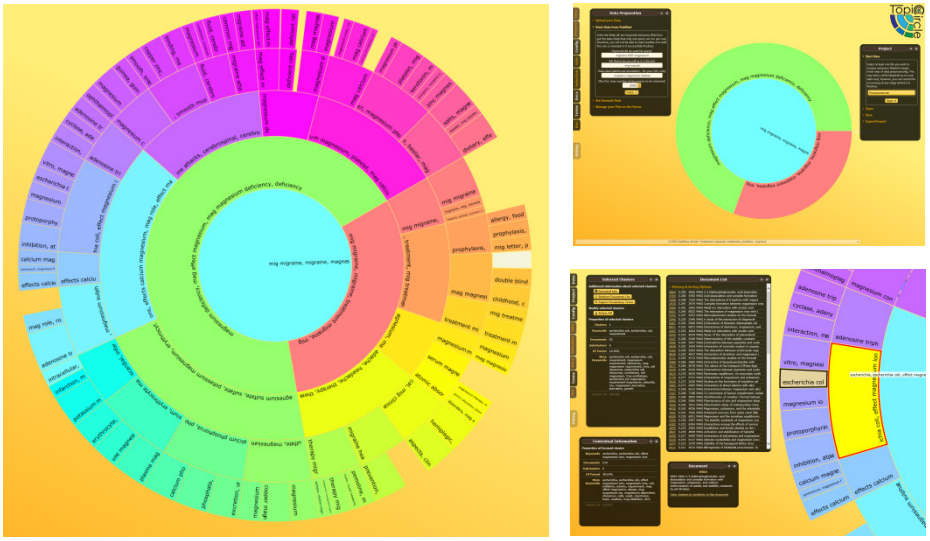


Fig. 4. The basic hierarchical cluster visualization is shown on the left along with two additional examples of screenshots of the application's data clustering functionalities on the right. The top-right screenshot presents data preprocessing and first splitting of documents to two sets (migraine and magnesium); the bottom-right screenshot presents zoomed-in view with a cluster selection, retrieved cluster documents and other contextual information.

Each of the clusters is described by its most meaningful keywords (written inside each cluster and displayed in detail when user moves the mouse over it). The bottom part of the right hand side of Fig. 4—the detailed information about a single document—shows one among many data representations, which support rich exploration, low threshold and wide walls as described by Schneiderman [17], [18].

The advantages of our new document cluster visualization, e.g., if compared to a well-established semi-automated cluster construction and visualization tool OntoGen [23], are that (a) the tool is not needed to download, (b) providing much more user friendly environment with especially low threshold for novice users to start exploring their data, and (c) providing wide walls with many different perspectives to the data—e.g. size of the cluster may be based on the number of sub clusters, included documents, or some other calculated property like similarity of the cluster to some query. Similarly is true for the color which may be used in a number of ways to help the user getting a better overview of the data. Fig. 5 presents an approach of using these properties (in this example we indeed use color) to better visualize cross-domain links which may be present in the data. When the user concentrates on a document in one domain he gets a suggestion of the similar clusters in both domains since all the similar clusters are emphasized with darker color. However, this is only one among many usages of the presented visualization for displaying additional rich cross-context aware information.

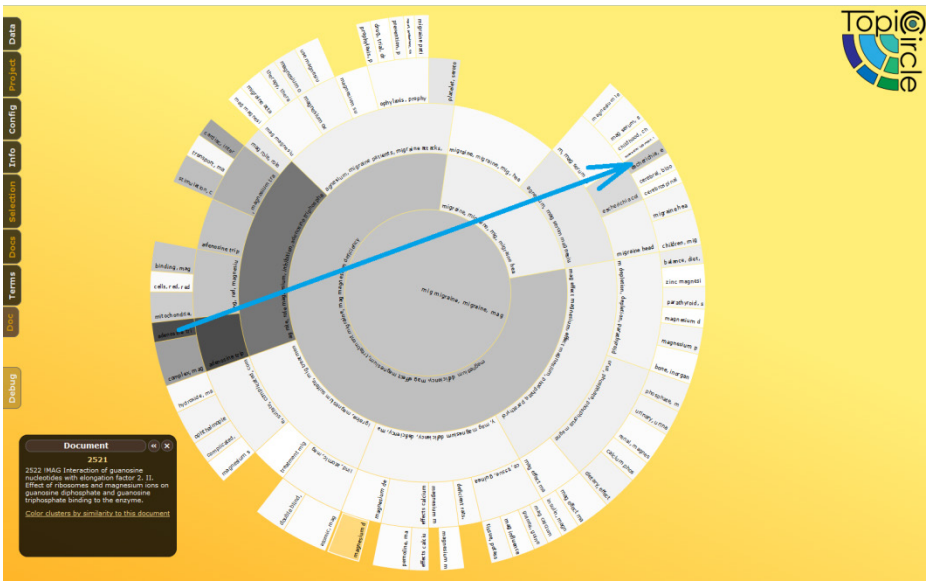


Fig. 5. Cluster colors can be used to show various information—in this case the cluster's similarity to a single selected document. The arrow shows similar clusters in two different domains, which can potentially indicate to a novel bisociative link between the two domains.

In terms of cross-context knowledge discovery, the top-down clustering approach enables the user to discover similar document sets within each domain, thus identifying potentially interesting domain subsets for further cross-domain link discovery using our system. Note that in the example presented in Fig. 4, clustering has been performed for each domain separately, therefore not fully demonstrating the potential for cross-domain knowledge discovery. In our past work, however, we have shown that when using clustering on a document set joining documents from both domains, the differences between the clusters identified using similarity measures using 2-means clustering and the document clusters identified through the initial document labeling by class labels of the two domains can fruitfully serve to identify outlier documents which include an increased number of B-terms and thus a high potential for B-term identification [24].

5 Conclusion

The paper presents a system for cross-context literature mining which supports experts in advanced document exploration by facilitating document retrieval, analysis and visualization. The system has been designed as a creativity support tool, helping experts in uncovering not yet discovered relations between different domains from large textual databases. As this is a very time-consuming process in which estimating linking potential of particular terms as well as efficient selection and presentation of pairs of documents to be inspected is very important, user interface has been designed very carefully. It supports experts by features such as visual document clustering,

color-based domain separation scheme, highlighted top-ranked bisociative terms and other functionalities, resulting in improved search capabilities needed for cross-domain discovery. A side-by-side view of documents from the two domains under investigation makes the discovery process easier in personal as well as in collaborative settings. Sessions with experts from different medical and biological domains have proved sophisticated functionality expected by experts. Together with a wide range of supplementary services the abovementioned characteristics contribute to the fact that the presented system can be viewed as a creativity support system.

The system and its user interface proved effective for cross-domain knowledge discovery in the two settings, described in the paper. However, heuristic user evaluation is out of the scope of the current paper and is left for further work. In particular regarding clustering, where the most important issue is the labeling particularly hard on higher levels, which are the more important levels from a navigational perspective [25], our work presented by Petrič et al. [24] already indicated the potential of using clustering for outlier document detection which are of ultimate importance for B-term identification.

In our future work we will introduce even more user interface options for data visualization and exploration as well as advance the term ranking methodology by adding new sophisticated heuristics which will take into account also more semantic aspects of the data. Besides, we will apply the system to new domain pairs to exhibit its generality, to investigate the need and possibilities of dealing with domain specific background knowledge, and last but not least to assist researchers in different disciplines on their way towards new scientific discoveries.

Acknowledgements. This work was supported by the Slovenian Research Agency and the FP7 European Commission projects MUSE (grant no. 296703) and FIRST (grant no. 257928).

References

1. Koestler, A.: The act of creation. MacMillan Company, New York (1964)
2. Agrawal, R., Mannila, H., Srikant, R., Toivonen, H., Verkamo, A. I.: Fast discovery of association rules. Advances in Knowledge Discovery and Data Mining. pp. 307–328 (1996)
3. Swanson, D.R.: Migraine and magnesium: Eleven neglected connections. Perspectives in Biology and Medicine. vol. 31, pp. 526–557 (1988)
4. Swanson, D. R.: Medical literature as a potential source of new knowledge. Bull Med Libr Assoc. vol. 78/1, pp. 29–37 (1990)
5. Lindsay, R. K., Gordon, M. D.: Literature-based discovery by lexical statistics. Journal of the American Society for Information Science and Technology. vol. 50/7, pp. 574–587 (1999)
6. Weeber, M., Vos, R., Klein, H., de Jong-van den Berg, L. T. W.: Using concepts in literature-based discovery: Simulating Swanson's Raynaud–fish oil and migraine–magnesium discoveries. J. Am. Soc. Inf. Sci. Tech. vol. 52/7, pp. 548–57 (2001)
7. Srinivasan, P.: Text Mining: Generating Hypotheses from MEDLINE. Journal of the American Society for Information Science and Technology. vol. 55/5, pp. 396–413 (2004)
8. Urbančič, T., Petrič, I., Cestnik, B.: RaJoLink: A Method for Finding Seeds of Future Discoveries in Nowadays Literature. In: Proceedings of the 18th International Symposium on

Foundations of Intelligent Systems. LNCS, vol. 5722, pp. 129–138. Springer, Heidelberg, Germany (2009)
9. Hristovski, D., Peterlin, B., Mitchell, J. A., Humphrey, S. M.: Using literature-based discovery to identify disease candidate genes. Int J Med Inform. vol. 74/2–4, pp. 289–298 (2005)
10. Yetisgen-Yildiz, M., Pratt, W.: Using statistical and knowledge-based approaches for literature-based discovery. J Biomed Inform. vol. 39/6, pp. 600–611 (2006)
11. Smalheiser, N. R., Swanson, D. R.: Using ARROWSMITH: a computer-assisted approach to formulating and assessing scientific hypotheses. Computer Methods and Programs in Biomedicine. vol. 57/3, pp. 149–153 (1998)
12. Holzinger, A., Yildirim, P., Geier, M., Simonic, K.-M.: Quality-based knowledge discovery from medical text on the Web Example of computational methods in Web intelligence. In: Advanced Techniques in Web Intelligence - Quality-based Information Retrieval Studies in computational intelligence. LNAI, vol. 452, pp. 145–148. Springer, Heidelberg, Germany (2013)
13. Dubitzky, W., Kötter, T., Berthold, M.R.: Towards creative information exploration based on Koestler's concept of bisociation. In: Bisociative Knowledge Discovery. LNCS, vol. 7250, pp. 11–32. Springer, Heidelberg, Germany (2012)
14. Juršič, M., Cestnik, B., Urbančič, T., Lavrač, N.: Bisociative Literature Mining by Ensemble Heuristics. In: Bisociative Knowledge Discovery. LNCS, vol. 7250, pp. 338–358. Springer, Heidelberg, Germany (2012)
15. 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)
16. Resnick, M., Myers, B., Nakakoji, K., Shneiderman, B., Pausch, R., Selker, T., Eisenberg, M.: Design Principles for Tools to Support Creative Thinking. In: Proceedings of the NSF Workshop on Creativity Support Tools. pp. 25–36 (2005)
17. Shneiderman, B.: Creativity support tools: accelerating discovery and innovation. Communications of the ACM. vol. 50/12, pp. 20–32 (2007)
18. Shneiderman, B.: Creativity Support Tools: A Grand Challenge for HCI Researchers. In: Engineering the User Interface. pp. 1–9. Springer, London, UK (2009)
19. Kranjc, J., Podpečan, V., Lavrač, N.: ClowdFlows: A cloud cased scientific workflow platform. In Proceedings: Machine Learning and Knowledge Discovery in Databases. LNCS, vol. 7524, pp. 816–819. Springer, Heidelberg, Germany (2012)
20. Urbančič, T., Petrič, I., Cestnik, B., Macedoni-Lukšič, M.: Literature Mining: Towards Better Understanding of Autism. In: Proceedings of the 11th Conference on Artificial Intelligence in Medicine. LNCS, vol. 4594, pp. 217–226. Springer, Heidelberg (2007)
21. Petrič, I., Urbančič, T., Cestnik, B., Macedoni-Lukšič, M.: Literature mining method RaJoLink for uncovering relations between biomedical concepts. Journal of Biomedical Informatics. vol. 42/2, pp. 219–227 (2009)
22. Salton, G., Buckley, C.: Term weighting approaches in automatic text retrieval. Inf. Process Manag. vol. 24/5, pp. 513–523 (1988)
23. Fortuna, B., Grobelnik, M., Mladenić, D.: Semi-automatic Data-driven Ontology Construction System. In: Proceedings of the 9th International Multiconference Information Society. pp. 212–220 (2006)
24. Petrič, I., Cestnik, B., Lavrač, N., Urbančič, T.: Outlier Detection in Cross-Context Link Discovery for Creative Literature Mining. The Computer Journal. vol. 55/1, pp. 47–61 (2012)
25. Muhr, M., Kern, R., Granitzer, M.: Analysis of structural relationships for hierarchical cluster labelling. In Proceeding of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 178–185 (2010)