Towards Narrative Ideation via Cross-Context Link Discovery Using Banded Matrices

Matic Perovšek[1,2], Bojan Cestnik[3,1], Tanja Urbančič[4,1], Simon Colton[5], and Nada Lavrač[1,2,4]

[1] Department of Knowledge Technologies, 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
[5] Department of Computing, Goldsmiths College, London, UK

Abstract. Knowledge discovery and computational creativity have until lately been investigated by two separate research communities. However, research in bisociative, cross-context knowledge discovery has recently started addressing creative tasks, including creative literature mining. This paper contributes to this effort by investigating an approach to cross-context link discovery based on banded matrices, aimed at identifying meaningful bridging terms (b-terms) at the intersection of two different domains. The proposed approach was applied to a simplified computational creativity task of narrative ideation from pairs of short sentences. As input, we took sentences from two different contexts: what-if sentences retrieved from Twitter, and morals from Aesop's fables, respectively. The approach resulted in a list of linked pairs of sentences from these two domains, illustrating the potential of the proposed approach to cross-context narrative ideation.

1 Introduction

The rate at which novelties are introduced in modern life has been so rapid that it invites changes in the way humans cope with creativity, especially how we gather information and how we make new connections between pieces of information. In the emerging field of computational creativity, Wiggins [1] has proposed the following definition: "computational creativity refers to performance of tasks (by a computer) which, if performed by a human, would be deemed creative".

This paper addresses a creative task of bisociatively connecting narratives from different domains. According to Koestler [2], a *bisociation* can be defined as (sets of) concepts that bridge two otherwise not (or very sparsely) connected domains. In his view, bisociative thinking occurs when a problem, idea, event or situation is perceived simultaneously in two or more different "matrices of thought" or domains. When two matrices of thought interact with each other, the result is either their fusion in a novel intellectual synthesis or their confrontation in a new aesthetic experience. Koestler regarded many different mental phenomena which are based on comparison (such as analogies, metaphors, jokes, identification, anthropomorphism, and so on), as special

A. Tucker et al. (Eds.): IDA 2013, LNCS 8207, pp. 333–344, 2013.

cases of bisociation. Following Koestler's ideas, the goal of this paper is the development of a computational system, which blends elements drawn from two previously unrelated matrices of thought into a new matrix of meaning.

In the area of knowledge discovery and literature mining, cross-domain connections have been the topic of recent research. In the area of bisociative knowledge discovery, Berthold [3] defines that two concepts are bisociated if there is no direct, obvious evidence linking them and if one has to cross different domains to find the link. Furthermore, a new link must provide some novel insight into the problem addressed. Several approaches to cross-domain knowledge discovery have been recently developed and reported [3]. In literature mining, on the other hand, cross-domain links in medical literature has been a topic of extensive research since early 1980s. For example, Smalheiser and Swanson [4] developed an online system ARROWSMITH which is supported by Swanson's ABC model approach. ARROWSMITH takes as an input two sets of titles from disjoint domains A and C and lists terms that are common to A and C. The resulting bridging terms (*b-terms*, forming set B) are further investigated for their potential to generate new scientific hypotheses.

Recently, the developers of the CrossBee system [5] investigated a specific form of bisociation, by exploring terms that appear in documents which represent bisociative links between concepts of different domains. Their methodology is based on an ensemble of specially tailored text mining heuristics which assign the candidate bridging concepts a bisociation score. The resulting ranked list of potential domain bridging terms enables the user to start inspecting b-terms with top-ranked terms, which should result in higher probability of finding observations that may lead to the discovery of new bridges between different domains. As described, the creative act is to find links which cross two or more different domains, leading out of the original "matrix of thought" or "out-of-the-plane" in Koestler's terms [2].

In our work we study heuristics for b-term ranking, resulting in an ordered list of potential bridging terms. The novel methodology introduced in this paper uses banded matrices [6] to discover structures which reveal the relations between the rows (representing documents) and columns (representing words/terms) of a given data matrix (representing a set of documents). We use this information in developing new heuristics for evaluating words/terms according to their bridging term (b-term) potential. In addition, the method enables the identification of document outliers, but this is out of the scope for this paper. While this methodology is mainly targeting cross-context knowledge discovery, this paper focuses on its use in the context of computational creativity for bisociative, cross-context narrative ideation. In the simplified narrative ideation scenario addressed in this paper, we use the proposed methodology for b-term ranking on documents from two domains to discover the bridging terms with the aim of combining sentences from the two domains. In our illustrative example, the first domain consists of 94 What-if sentences taken from Twitter, while the second domain consists of 277 morals from Aesop's fables. The presented approach resulted in a list of interesting linked pairs of sentences from both contexts. An example of such a pair of sentences is: *"What if humans could actually breathe in space and the government says we can't so we don't try to escape? Nothing escapes the master's eye."*

The paper is structured as follows. In the next two sections, we describe the methodology of discovering links between two unrelated contexts by using banded matrices. In the experimental section, we present the results of our methodology in a simplified narrative ideation scenario, to combine statements from two different domains. We conclude the paper by highlighting the most important findings and providing some plans for further work.

2 Banded Matrices: Definition and a Motivating Example

Our research aims at finding cross-domain links by exploring the bridging terms (b-terms) at the intersection of domains that establish links between two domains of interest. The proposed approach follows the work of Juršič et al. [5], which already contributed to b-term detection in cross-context literature mining. The approach to cross-context link discovery that we investigate in this paper is new: it is based on banded matrices [6].

Our methodology works by first encoding the documents from the two domains into the standard Bag-Of-Words (BOW) vector representation and then transforming the binary matrix of BOW vectors to its banded structure. The proposed banded matrices methodology is based on the assumption that similar documents, as well as the words that appear in the same document, will appear closer to each other in the matrix and will therefore form "clusters" along the main diagonal of the matrix in its banded form[1]. Our work is based on the intuition that terms that connect different domains will be positioned at the edges of clusters from different domains: we have developed different heuristics that should be able to identify these b-terms by ranking them high in the ranked list of terms with high potential for cross-context link discovery. We introduce below the banded matrices, and follow this with a toy example illustrating the approach.

2.1 Definition of Banded Matrices

Uncovering structures that reveal the nature of relations between rows and columns of data matrices is an important step towards solving real-world problems, as binary data occur in numerous real-world applications. Recent research in social networks, bioinformatics, and human genomics has shown the benefits of banded representations of matrices [7]. These representations have contributed to bringing huge performance boosts in various mathematical operations, including matrix multiplication.

To explain the algorithm that transforms a matrix into its banded structure we first need to define the basic concepts. A binary matrix has a banded structure if we can find a permutation of rows and columns such that the 1s exhibit a staircase pattern down the rows along the leading diagonal, as illustrated in Figure 1.

A binary matrix M is fully banded if there exists a permutation of rows κ and a permutation of columns π such that:

1. *for every row i in M_κ^π the entries with 1s occur in consecutive column indices $a_i, a_i + 1, \ldots, b_i$, and*
2. *the values of starting indices for 1s in successive rows (i and i+1) satisfy $a_i \leq a_{i+1}$ and $b_i \leq b_{i+1}$.*

[1] A correspondence between bi-clustering and banded structures has been shown in [7].

A necessary precondition for (1) to hold is that matrix M satisfies the *consecutive-ones property*: a binary matrix satisfies this property if it is possible to order the columns so that in every row the non-zero entries occur in consecutive positions.

As banded structured matrices cannot be expected to arise in noisy real-world environments, we need to redefine the problem in the sense that it is applicable to a wider range of real-world situations. We aim to minimize the number of transformations one needs to perform on a binary matrix to unveil a banded structure. The number of such transformations will measure how far the matrix is from being fully banded. The algorithm presented in the next section (following the motivating example) aims to solve this optimization problem.

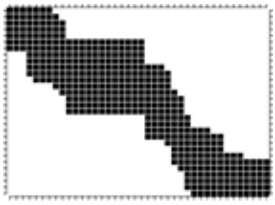

Fig. 1. An example of a fully banded matrix

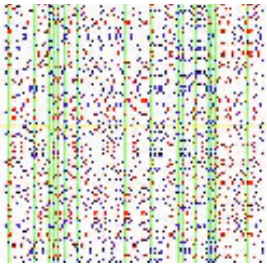

Fig. 2. Documents (rows) and words (columns) in an ideal-world domain. The color of the row indicates the domain of the document (blue for domain A and red for domain C).

2.2 A Motivating Example

Let us have two sets of documents A and C. For the purpose of explaining the methodology we constructed a small ideal-world dataset, which consists of 6 clusters of documents, 3 of which belong to domain *A*, while the others belong to domain *C*. Initially, we took a set of 120 different randomly selected words and randomly divided them into 6 clusters, so that there were no intersections, i.e., each word belonged to one cluster only. The number of possible words per cluster was 20, while each document in the cluster was randomly assigned only 15 of these words. Using a banded matrix algorithm presented in the next section this document set would first be transformed into the structure shown in Figure 3 and finally into a fully banded matrix form shown in Figure 4.

In order to illustrate our methodology, we randomly chose 8 words from each of the two domains A and C to act as artificially defined, preselected bridging terms. This effect was achieved by inserting the preselected terms into every document in every cluster with a 10% chance, thus spoiling the original clean separation of words within documents of different clusters. The resulting matrix showing documents as rows, and words as columns, is depicted in Figure 2, where the green vertical lines represent the artificially inserted b-terms. As the aim of our method is to identify the bridging terms,

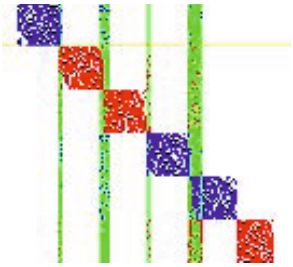

Fig. 3. Final result of our methodology: matrix of documents from Figure 2 permuted using row and column permutations obtained from the transformation of the matrix into its fully banded structure

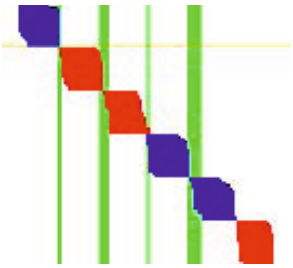

Fig. 4. Matrix of documents shown after the transformation of documents in Figure 2 into a fully banded matrix structure. Rows represent the documents, while columns represent the terms. The green vertical lines represent the terms which were inserted as potential bridging terms to the documents.

we conducted experiments to check how the designed preselected terms will be ranked by our heuristics.

Having used our methodology on the ideal-world toy domain of Figure 2, we got the result shown in Figure 3. The green vertical lines represent the terms which were deliberately acting as bridging terms in this experiment. As can be seen from Figure 3, similar documents (documents from the same cluster) and similar terms (terms that are contained in the same document cluster) are located close to each other. As a result, the "clusters" along the matrix leading diagonal are clearly visible. Note that the preselected bridging terms occur mainly on the transitions between the clusters. All of our heuristics (explained in the next section) correctly assigned a b-term score greater than 0 only to the preselected bridging terms, which served as a proof-of-concept for the toy experiment.

Let us now consider a single document only. A document from domain A (represented with a horizontal yellow line on Figures 2, 3 and 4 consist of the following words: *magnesium blood cell prophylaxi relationship lithium red calcium effect sodium control membrane measurement potential perfuse* ***ophthalmoplegic simultaneous***, where the first 15 words were randomly selected from the document's cluster term set, while the two words in bold were randomly inserted from the preselected set of bridging terms. The blue dots on of the horizontal yellow line in Figure 3 consequently symbolize the above words. According to the banded structure of the matrix (see Figure 4) the words ***simultaneous*** and ***ophthalmoplegic*** belong to word clusters of domains A and C, respectively. While the observed document belongs to domain A, the term ***ophthalmoplegic*** is representative for the documents from domain C. Therefore, our methodology should be able to identify this term as a potential b-term. In contrast, as the word ***simultaneous*** is used in the documents from the same domain A, it should not be considered

as a b-term. Indeed, our heuristics (presented in the next section) have only identified ***ophthalmoplegic*** as a b-term. Figure 4 shows the final result of the banded matrix algorithm and will be used in the next section for the explanation of our heuristics.

3 A Methodology for B-Term Ranking Using Banded Matrices

Our approach is designed to find links between two domains, named A and C, by exploring the bridging terms that connect these two separate domains. The methodology works as follows. First, we preprocess the documents from the two domains using standard text mining techniques [8]. This is performed through a number of steps: stop-word removal, stemming or lemmatization, usage of synonym dictionaries, construction of n-grams of words and, finally, transformation to a Bag-Of-Words representation. Next, the result of the preprocessing step, i.e., the binary matrix of "Bag-Of-Words" vectors (the BOW matrix), is transformed to its banded matrix structure. Finally, we sort the terms according to their scores representing their bridging term potential, computed according to the new heuristics described below. In the following subsections, each step of the proposed bridging term detection and ranking methodology is described in detail.

3.1 Constructing a Banded Structure Using a Bidirectional MBA Algorithm

The optimization problem addressed to make a banded matrix is labeled Bidirectional Minimum Banded Augmentation (bidirectional MBA) [6] and is defined as follows:

Given a binary matrix M, find the minimum number of bidirectional flips (flips from both 1s to 0s and 0s to 1s) so that M becomes fully banded.

Algorithm 1: Bidirectional MBA algorithm

1. Find fixed permutation of columns π.
2. Solve the consecutive-ones property on the column permuted matrix M^{π}.
3. Resolve Sperner conflicts (defined later in this section) between rows in M^{π}.
4. Sort the rows in M^{π} and return the row permutation.

The presented MBA algorithm discovers a single band by first fixing the column permutations of the data matrix before proceeding with the rest of the algorithm. A good permutation of columns tends to put similar columns (i.e., terms) closer to each other. We use the Jaccard coefficient as a column similarity measure: $J(M^a, M^b) = \frac{M^a \cap M^b}{M^a \cup M^b}$. In our example, this similarity measure returns the highest value of 1 when two terms occur in the same set of documents. We used the spectral ordering algorithm [9] to find the fixed column permutation π of matrix M.

Next, the algorithm deals with solving the consecutive-ones properties on rows of matrix M^{π}, which is an essential step in finding the row permutation κ. Solving the consecutive-ones problem for row M_i^{π} with bidirectional flips corresponds to solving the maximum sub-array problem on matrix W_i^j [6], defined as follows:

$$W_i^j := \begin{cases} 1 & \text{if } M_i^j = 1 \\ -1 & \text{if } M_i^j = 0 \end{cases}.$$

The objective of solving the maximum sub-array problem is to find the sub-array of the matrix that has the maximum sum of numbers. This problem can be solved in linear time with respect to the size of the array using the scan-line algorithm [10]. This method returns the interval boundaries which we use to solve the consecutive-ones problem in M_i^π by setting the fields in M_i^π between the boundaries to 1 and others to 0.

Next, the algorithm deals with removing the Sperner conflicts between the rows of matrix M^π. A matrix has Sperner conflicts if its rows do not form a Sperner family of intervals:

Two rows $M_i = [a, b]$ and $M_j = [a', b']$ with consecutive-ones property, where $i \neq j$, form a Sperner family of intervals if they are overlapping such that $(a \geq a' \vee b' \geq b) \wedge (a' \geq a \vee b \geq b')$.

Additional flips on rows of M^π need to be made in order to ensure that rows have the Sperner family of intervals property.

Let $\hat{M}$ be the binary matrix M augmented with $M_{ij} = M_i \backslash M_j$ for every two rows $M_i \subset M_j$. Note that M is fully banded if and only if $\hat{M}$ has the consecutive-ones property (proof in [6]).

To eliminate all Sperner conflicts between row intervals of M^π, the algorithm has to go through all extra rows described in $\hat{M}$ and make sure that they have the consecutive-ones property. This can be done by solving the maximum sub-array problem on the extra rows of $\hat{M}$. We perform additional flips in order for the rows to obtain consecutive-ones property. Lastly, we update the rows in M^π according to the changes made over $\hat{M}$ in order to get a banded matrix.

Finally, the algorithm sorts rows $[a, b]$ of M^π in an ascending order of as, while deciding ties with the ascending order of their bs. The result of the algorithm is a banded matrix M_{band}, along with details of the row and column permutations that were performed. We use these permutations on our original matrix M, as the objective is to produce a matrix without distorting the data (i.e. without making the bidirectional flips). In the next section, we present the heuristics for calculating the b-terms potential scores.

3.2 New Heuristics for B-Terms Potential Evaluation

Here we describe the details of the four heuristics which we propose for computing the bridging term potential scores. After completing the step of term score computation, we sort the terms according to the values of one of the heuristics and present the top-ranked terms (hopefully representing the most interesting b-term candidates) to the expert. The designed heuristics should favor b-terms over non-b-terms by pushing interesting b-term candidates higher to the top of the ranked term list. For easier definition of the proposed heuristics we define variable d_{idx} to represent the row index of document d in the banded matrix M_{band} and t_{idx} to represent the column index of term t in M_{band}. Note that in order to compute the score of the proposed heuristics, we distinctively take

into account the document-term matrix in two forms, banded (as shown in Figure 4) and full (as shown in Figure 3).

Heuristic 1: This is a frequency based heuristic for computing the b-term potential. If all document-term pairs in the t_{idx}-th column of matrix M_{Band}, which equal to 1, belong to the same domain, we denote this domain as D_1. Note that in such a case, the t_{idx}-th column, which represents term t in the banded matrix, should be "single-colored" in the matrix visualization in Figure 4. If the documents in the M_{Band} for term t do not belong to the same domain, the heuristic returns score 0 for this term ($h1score(t) := 0$). Otherwise, the score of Heuristic 1 for term t is defined as:

$$h1score(t) := countDoc_{D_2}(t),$$

where $countDoc_{D_2}(t)$ is the number of documents that contain term t and belong to domain D_2 (do not belong to domain D_1) in the matrix shown in Figure 3. This heuristic is based on the assumption that terms which strongly represent one domain (the single-colored column in the banded matrix of Figure 4), and at the same time there are many documents from the other domain that contain these terms, have a higher chance of being the bridging terms between the two domains.

Heuristic 2: This is also a frequency based heuristic. Similarly as described in Heuristic 1, if all documents for which the t_{idx}-th column of matrix M_{Band} equals to 1 belong to the same domain, we label this domain as D_1. Otherwise, the heuristic returns score 0 for term t ($h2score(t) := 0$). Heuristic 2 score for term t is defined as:

$$h2score(t) = \frac{countDoc_{D_2}(t)}{countOnDiagDoc_{D_1}(t)},$$

where $countDoc_{D_2}(t)$ is the number of documents from domain D_2 that contain term t, and $countOnDiagDoc_{D_1}(t)$ is the count of document-term pairs equaling 1 in the t_{idx}-th column of the banded matrix M_{Band}: $countOnDiagDoc_{D_1}(t) := |\{d_{idx}; d \in D_1 \wedge M_{Band}(d_{idx}, t_{idx}) = 1\}|$. Therefore, for term t $h2score(t)$ is the ratio between the count of documents that belong to domain D_2 and the documents in the M_{Band} for term t which belong to domain D_1. The intuition behind this heuristic is that a term that strongly represents a given domain according to the banded matrix, and is at the same time also contained in many documents of the other domain, should also have a high b-term potential score.

Heuristic 3: This is an inverse of Heuristic 2, defined as:

$$h3score(t) := \begin{cases} 0 & \text{if } h2score(t) = 0 \\ \frac{1}{h2score(t)} & \text{otherwise} \end{cases}$$

The intuition behind this heuristic is that a term that strongly represents a domain in banded matrix M_{band} should get a higher score compared to the other terms. The number of documents on the diagonal should be as high as possible: column t_{idx} should contain as many document-term pairs from the same domain.

Heuristic 4: If the documents in M_{Band} for term t do not belong to the same domain, this heuristic returns a score of 0 ($h2score := 0$). Otherwise, we label this domain as D_1 and define the Heuristic 4 score for term t as follows:

$$h4score(t) = \frac{countOnDiagDoc_{D_1}(t)}{countDoc_{D_1}(t)} * countDoc_{D_2}(t),$$

where $countOnDiagDoc_{D_1}(t)$ is the count of document-term pairs in the t_{idx}-th column of banded matrix M_{Band}, and where documents belong to domain D_1: $countOnDiagDoc_{D_1}(t) = |\{d_{idx}; d \in D_1 \wedge M_{Band}(d_{idx}, t_{idx}) = 1\}|$; $countDoc_{D_1}(t)$ denotes the number of documents from domain D_1 that contain term t, while $countDoc_{D_2}(t)$ is the number of documents from domain D_2 that contain term t. The bridging term potential score for term t is the ratio of documents from domain D_1 that lay on the diagonal multiplied by the number of documents from the other domain. The intuition behind this heuristic is that for term t, the more the term represents a domain (has a large proportion of document on the diagonal of the banded matrix) and also the more documents from the other domain that contain t exist, the higher the potential of term t to be a bridging term between the two domains.

All the heuristic scores are normalized by dividing the term scores with the highest score. The result of our methodology is a list of terms sorted by their b-term potential scores. It should be left to the domain expert to check whether the discovered bridging term suggests a valid, new and interesting relation. While the methodology has been currently applied to bridging two domains only, the method can be generalized to connecting several different domains, which is the topic of our further research.

4 Using Banded Matrix-Based B-Term Ranking for Bisociative Morals Ideation

Although not primarily designed for this task, our methodology can be used for creating pairs of sentences from different domains, which combine into surprising, funny or even insightful pieces of text when put together and considered as a whole. Bridging terms, appearing in both sentences, detected by our methodology function as a kind of glue, contributing to the coherency and increasing the potential for combinations to be meaningful.

To illustrate the potential of the proposed approach for narrative ideation, we chose two domains: What-if sentences and Aesop's fables' morals. The first domain consists of 94 What-if sentences, retrieved from Twitter (using hash tag #whatif) and the UKWaC British English web corpus. For instance, here is example of such a sentence: *"What if Google someday went down and we couldn't Google what happened to Google?"* The Aesop's fables[2] morals dataset is a collection of 277 fable morals. We created this dataset by crawling the Aesop's fables online collection.

[2] Aesop was a Greek fabulist, known as author of numerous fables; these are characterized by animals, which solve problems and have human characteristics. See `http://www.aesopfables.com/`

Table 1. Results of our methodology: combinations of what-if sentences with Aesop's fable morals. In the brackets are the fable titles, which were not part of the document's contents, but are given here as an additional piece of information. The bridging terms are shown in bold.

What if **life** is one big dream, and when we die, we wake up. Evil tendencies are shown in early life. (The Man, the Boy and the Donkey)
What if we woke up, as a baby, and our whole life had been a dream? Evil tendencies are shown in early **life**. (The Man, the Boy and the Donkey)
What if, like Bhutan, we gauged our **life**'s success by how happy we are, not by how big the house is, the number. Evil tendencies are shown in early **life**. (The Man, the Boy and the Donkey)
What if someone you **love** dearly gave you a surprise bday party and when you arrived every1 was exchanging gifts but not nary a 1 was for you! Misery **loves** company. (The Fox Who Had Lost His Tail)
What if someone you **love** dearly gave you a surprise bday party and when you arrived every1 was exchanging gifts but not nary a 1 was for you! Even the wildest can be tamed by **love**. (The Lion in Love)
What if there are other beings in the same room as us but we can't **see** them and they can't see us. Not everything you **see** is what it appears to be. (The Dancing Monkeys)
What if there are other beings in the same room as us but we can't **see** them and they can't see us. Gossips are to be **seen** and not heard. (The Eagle the Cat and the Wild Sow)
What if we don't **see** tomorrow and everything you said today couldn't be undone. Would you be proud? Not everything you **see** is what it appears to be. (The Dancing Monkeys)
What if we don't **see** tomorrow and everything you said today couldn't be undone. Would you be proud? Gossips are to be **seen** and not heard. (The Eagle the Cat and the Wild Sow)
What if eyeballs had butts but we couldn't **see** them because they're hidden in our skulls? Not everything you **see** is what it appears to be. (The Dancing Monkeys)
What if eyeballs had butts but we couldn't **see** them because they're hidden in our skulls? Gossips are to be **seen** and not heard. (The Eagle the Cat and the Wild Sow)
What if humans could actually breathe in space? And the government says we can't so we don't try to **escape**? Nothing **escapes** the master's eye. (The Stag in the Ox-Stall)
What if humans could actually breathe in space? And the government says we can't so we don't try to **escape**? We had better bear our troubles bravely than try to **escape** them. (The Kings Son and the Painted Lion)
What if you simply stopped doing whatever it is that isn't a part of your **success**? The best intentions will not always ensure **success**. (The Monkeys and Their Mother)
What if, like Bhutan, we gauged our life's **success** by how happy we are, not by how big the house is, the number. The best intentions will not always ensure **success**. (The Monkeys and Their Mother)

Each what-if sentence as well as each Aesop's fables' moral was treated as a separate document. The documents from both domains were preprocessed using standard text mining techniques, described in the methodology section. This resulted in 383 distinct terms from all the obtained documents. We applied our methodology to find terms with the highest b-term potential. For simplicity, the terms were sorted using the scores of Heuristic 4, which we consider the most complete among the presented heuristics. The five terms with the highest b-term potentials were used to create all pairs of sentences sharing the selected bridging term (with the first sentence being from the what-if domain and the second from the Aesop's fables domain). The 15 highest scoring b-term based

concatenated pairs of sentences which resulted are shown in Table 1. These results show–subjectively–that using the terms with the highest b-term potential resulted in several meaningful, creative combinations of sentences. Moreover, it is clear that a large proportion of the sentence pairs in Table 1 have meaningful relations which could form the basis of artefacts such as poems or stories. We can argue that it would be quite a laborious task to find similarly valuable combinations from all possible pairs (143) of sentences from the given domains without guidance provided by bridging terms. We plan to use crowd-sourcing to test the hypothesis that our approach can reliably produce such valuable combinations.

5 Conclusions and Further Work

The experimental evidence above indicates that the methodology presented in this paper has the potential for supporting the users in the task of bisociative, cross-domain narrative ideation. Banded matrices help us to discover the structures which reveal the nature of relations between terms and documents. We have shown that the approach can be used to construct creative combinations of sentences from different domains, coupled with bridging terms with the highest b-term potential. The results confirm the potential of the proposed approach to identify meaningful bridging concepts in the intersection of texts from different domains.

In further work, we will upgrade the methodology to combine not only pairs of sentences from two domains, but also to compose longer chains of sentences, resulting in narrative ideation. Another line of research will address more subtle connections between sentences. For instance, there is a semantic connection between *baby* and *early life* in *"What if we woke up, as a baby, and our whole life had been a dream? Evil tendencies are shown in early life."* However, this connection arises purely by coincidence: it was not detected by the system. Furthermore, introducing more semantic understanding into the ranking could substantially improve the performance. We further plan to improve to the set of heuristics, e.g., by introducing a more global view taking into consideration a term's local neighbourhood, and exploring the potential of outlier documents in guiding search. We plan to apply this narrative ideation approach to knowledge discovery in different domains, working with experts from different fields to address real-life artistic and scientific domains and getting valuable feedback from them.

An important next step will be to crowd-source opinions about how reliable the process is at producing sentence pairs (and larger constructs) which can be meaningfully interpreted in such a way that intelligence and possibly creativity are projected onto the software producing them. After the analysis of the results and further refinement of the techniques, we plan to embed fictional ideation processes and idea expansion via bisociation into software for generating artefacts of cultural value such as poems, stories and scientific hypotheses. We hope to show that the kinds of cross-context link discovery methods presented here can be used generically in Computational Creativity projects across domains.

References

1. Wiggins, G.: A preliminary framework for description, analysis and comparison of creative systems. Knowledge-Based Systems 19(7), 449–458 (2006)
2. Koestler, A.: The Act of Creation, vol. 13 (1964)
3. Berthold, M. (ed.): Bisociative Knowledge Discovery. Springer (2012)
4. Smalheiser, N., Swanson, D., et al.: Using ARROWSMITH: A computer-assisted approach to formulating and assessing scientific hypotheses. Computer Methods and Programs in Biomedicine 57(3), 149–154 (1998)
5. 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, pp. 33–40 (2012)
6. Garriga, G., Junttila, E., Mannila, H.: Banded structure in binary matrices. Knowledge and Information Systems 28(1), 197–226 (2011)
7. Alqadah, F., Bhatnagar, R., Jegga, A.: Mining maximally banded matrices in binary data. In: Proceedings of the 10th SIAM International Conference on Data Mining (SDM 2010), pp. 942–953 (2010)
8. Feldman, R., Sanger, J.: The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data. Cambridge University Press (2006)
9. Atkins, J., Boman, E., Hendrickson, B.: A spectral algorithm for seriation and the consecutive ones problem. SIAM Journal on Computing 28(1), 297–310 (1998)
10. Cormen, T., Leiserson, C., Rivest, R., Stein, C.: Introduction to Algorithms. MIT Press (2001)