Bridging term discovery for cross-domain literature mining

Matic Perovšek[1,2], Nada Lavrač[1,2,3], and Bojan Cestnik[4,1]

[1] Department of Knowledge Technologies, Jožef Stefan Institute, Ljubljana, Slovenia
[2] International Postgraduate School Jožef Stefan, Ljubljana, Slovenia
[3] University of Nova Gorica, Nova Gorica, Slovenia
[4] Temida d.o.o., Ljubljana, Slovenia

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. This kind of context crossing associations, named bisociations [1], are often needed for creative, innovative discoveries. Following Koestler's ideas [1] and based on computational approaches to bisociative knowledge discovery [2], the goal of this work is to develop a computational system able to discover links between two previously unrelated domains, represented by two different document corpora. The work upgrades the CrossBee methodology [3], aimed to detect the bridging terms that represent bisociative links between different domains. The CrossBee methodology employs an ensemble of specially tailored text mining heuristics that assign to each discovered candidate bridging term (B-term) a score, which should reflect their bisociation potential. The resulting ranked list of potential B-terms enables the user to inspect the top-ranked B-terms, which may result in higher probability of finding observations that lead to the discovery of new bridges between the two domains.

We have extended the CrossBee system by integrating a new complementary technique to B-term ranking. The proposed technique uses banded matrices [4] to discover structures which reveal the relations between the rows (representing documents) and columns (representing terms) of a given data matrix representing a set of documents. We use this information in computing new heuristics that evaluate terms according to their potential for B-term discovery. The proposed approach thus encodes the documents from the two domains into the standard Bag-Of-Words (BOW) vector representation and then transforms the binary matrix of BOW vectors into a banded structure.

The proposed banded matrix methodology is based on the assumption that similar documents, as well as the terms 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. Our work is based on the intuition that terms that connect different domains will be positioned at the edges of clusters from different domains, and the developed heuristics should be able to identify these B-terms by ranking them high in the ranked list of terms with high potential for cross-domain link discovery.

The methodology we propose works as follows: first, we preprocess the documents from the two domains using standard text mining techniques. 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 into the banded matrix structure. In the next step we permute columns and rows of the binary matrix using

a bidirectional MBA algorithm [4] in order to retrieve a banded structure, followed by using the proposed heuristics to calculate the B-term potential for every term. One of the proposed heuristics calculates this score for term t as the ratio of documents characterised by term t in a document cluster grouped around the diagonal, multiplied by the number of documents from the other domain including term t. 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 cluster 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. After completing the step of term score computation, we sort the terms according to the values 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 to the top of the ranked term list.

We are mostly interested in the quality of heuristics from the end-user's perspective. Note that the standard ROC curves and AUC statistics do not provide the most significant evidence of the quality of individual heuristics, even though—in general—a better ROC curve reflects a better heuristic. Usually the user is interested in questions like: how many B-terms are likely to be found among the first n terms in a ranked list (where n is a selected number of terms the expert is willing to inspect, e.g., 5, 20 or 100).

In the experiments we used the well-researched migraine-magnesium domain pair as introduced by Swanson [5]. Swanson managed to find more than 60 pairs of articles connecting the migraine domain with the magnesium deficiency via 43 bridging concepts (B-terms). Using the developed methodology we tried to rank these 43 B-terms as high as possible among other terms that are not marked as B-terms. It proves that banded matrices help us discover the structures that indeed reveal the relation between terms and documents, which allows for faster cross-domain discovery than with the original CrossBee tool. Furthermore, we show that by using a predefined vocabulary we can increase the heuristic's capacities to rank the B-terms at the beginning of the term list. Indeed, by applying this approach in the migraine-magnesium domain we got a higher concentration of Swanson's B-terms among the best ranked terms. Consequently, the user is presented with a simpler exploration task, potentially leading to new discoveries.

Acknowledgment. *This work was supported by the Slovenian Research Agency grant as well as the FP7 European Commission projects MUSE (grant agreement no: 296703) and ConCreTe (grant agreement no: 611733).*

References

1. Koestler, A.: The Act of Creation. Volume 13. (1964)
2. Berthold, M., ed.: Bisociative Knowledge Discovery. Springer (2012)
3. 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) 33–40
4. Garriga, G., Junttila, E., Mannila, H.: Banded structure in binary matrices. Knowledge and Information Systems **28**(1) (2011) 197–226
5. Swanson, D.R.: Migraine and magnesium: Eleven neglected connections. Perspectives in Biology and Medicine **78**(1) (1988) 526–557