Text Mining for Discovering Implicit Relationships in Biomedical Literature

Ingrid Petrič
University of Nova Gorica, Vipavska 13, 5000 Nova Gorica, Slovenia
ingrid.petric@ung.si, http://slais.ijs.si/theses/2009-10-30-Petric.pdf

Thesis Summary

Keywords: text mining, knowledge discovery, hypotheses generation

Received: April 15, 2009

This article presents an innovative methodology for knowledge discovery in text databases that can improve the existing methods of exploring implicit relationships across different domains of expertise by providing a more intuitive computer aided search of unexplored links in literature. The literature mining method, called RaJoLink is based on rare pieces of information in a given domain. When these relations are interesting from a medical point of view and can be verified by medical experts, they represent new pieces of knowledge and can contribute to better understanding of diseases.

Povzetek: Članek opisuje inovativno metodologijo odkrivanja znanja iz tekstovnih baz podatkov.

1 Introduction

Automated knowledge discovery based on text data sets in the field of biomedicine is an intriguing challenge as it requires intensive collaboration with domain experts during the processes of both domain-specific text analysis and evaluation. Hence an interactive approach is recommended when text mining and decision support are combined. Also, it is beneficial to apply improved methods of literature mining, searching indirect connections and bisociative knowledge discovery from extensive text databases such as MEDLINE. Namely, information that is related across different contexts is difficult to identify with conventional associative approaches. The context-crossing associations, however, are the ones often needed for innovative discoveries. Such associations are called bisociations. The major aim here is to unravel the still hidden relations between the researched phenomena and their potential causes.

We developed a literature mining method called RaJoLink [1-3, 6] that uncovers hidden relations from large sets of scientific articles in a given domain. The method searches for logically connected pieces of literature on rare terms identified in literature on a given phenomenon under investigation (e.g., disease). This way it supports human expert in the process of generating and testing hypotheses in the domain under study. RaJoLink supports biomedical experts in both open and closed discovery process. In the open knowledge discovery process, hypotheses have to be generated, while in the closed knowledge discovery process, given hypotheses are tested. By identifying relations between biomedical concepts in disjoint sets of articles, the method implements the Swanson's [4] ABC model approach. However, the RaJoLink method analyses such relations in a new way and expands the Swanson's ABC model by suggesting hypotheses in advance, as a result of the open knowledge discovery process. The main novelty is a semi-automated suggestion of candidates for agents *A* that might be logically connected with a given phenomenon *C* under investigation. The choice of candidates for *A* is based on rare terms identified in the literature on the topic *C*. As rare terms are not part of the typical range of information, which describe the phenomenon under investigation, such information might be considered as unusual observations about the phenomenon *C*. If literatures on these rare terms have an interesting term in common, this joint term is declared as a candidate for *A*. Linking terms *B* between literature on *A* and literature on *C* are then searched for in the closed discovery process to provide supportive evidence for uncovered connections.

2 RaJoLink

The method is named RaJoLink after its key procedural elements, which are: rare terms, joint terms and linking terms. Consequently, the entire RaJoLink's approach consists of three principal steps, Ra, Jo and Link. In step Ra, a specified number (set by user as a parameter value) of 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. One of them is selected as the candidate for *A*. In step Link, linking terms *B*, which bridge literature about *A* and literature about *C*, are searched for. Relations between *A* and *C* are established via *AB* and *BC* relations. Evaluation of pairs (*AB*, B*C*) as support for

potential hypotheses about the relation between *A* and *C* is carried out by the domain expert.

We have applied the RaJoLink method to the scientific literature on autism and have used MEDLINE as a source of data. Autism was selected as the problem domain due to its complexity, insufficient and partial knowledge about its various causes, and because of the strong focus of current medical research towards early diagnosis of this disorder. In the autism domain we discovered a relation between autism and calcineurin and between autism and transcription factor NF-kappaB, which have been evaluated by a medical expert as relevant for better understanding of autism [3]. To assess the usefulness of RaJoLink in general, we evaluated the potential of our method also in the migraine-magnesium experiment, which represents a gold standard for the literature-based discovery [5]. For all these purposes we also developed a software tool, which implements the RaJoLink method and provides decision support to experts in the process of generating and testing of the scientific hypotheses in biomedical domains.

The unique contribution of the RaJoLink method and a fundamental difference from the previously proposed models of the literature-based knowledge discovery approach lies in the rarity principle that we apply to the open literature-based discovery. In fact, we use rare terms identified in the literature *C* to guide the search for new hypotheses. To this end, we have applied the rarity principle together with the notion of bisociation. In fact, the context-crossing connections, called bisociations, are often needed for creative, innovative discoveries. Bisociative relationships can only be discovered on the basis of a sufficiently large and diverse underlying corpus of information. In our case this corpus are MEDLINE papers. The larger the corpus is, the more likely it is to contain bisociative relationships. The RaJoLink approach has the potential for bisociative relation discovery as it allows switching between contexts (papers from different areas) by exploring rare terms in the intersection between contexts.

Besides this, we contributed an innovative approach also to the closed discovery process. The closed discovery process in RaJoLink is based on the outliers' detection in the content similarity graphs. In fact, having two disjoint literatures *A* and *C*, we automatically search for linking terms that are mentioned in both, the literature *A* as well as in the literature *C*. Pairs of documents with the same linking terms are subject to closer inspection in order to find out whether by putting statements about a linking term in these two articles together supports the hypothesis about a meaningful relation between previously disjoint literatures. In this manner, our search for linking terms is done in a semi-automated way that reduces manual work and efficiently points to meaningful relations between the domains *A* and *C*. In the closed discovery process we presented important connections between autism and calcineurin literature, as well as between autism and NF-kappaB literature. We discovered such connections by analysing outliers in the published evidence of some autism findings on one hand that coincide with specific calcineurin and NF-kappaB observations on the other hand.

3 Conclusion

The knowledge gathered by various specialised sciences throughout the digital era has resulted in large volumes of data and complex data interrelationships. To support biomedical experts in their knowledge discovery process, we have developed a literature mining method called RaJoLink. One of the main advantages of RaJoLink lays in the support of knowledge discovery by the innovative use of rare terms from the problem domain literature to guide the generation of new hypotheses. Accordingly, the crucial step of the method consists of selecting rare terms from the problem domain literature. The intuition behind this research idea was that the rarer a term is in the domain literature, the higher is the probability to encounter observations that represent something unexpected that may lead to creative discovery of new knowledge. This way we managed to employ rarity as a principle and means to find new interesting pieces of knowledge that were previously available in the dispersed literature and could be linked together. The results of the experimental case studies in autism [1-3, 6] and migraine [5] domain showed that the RaJoLink method can enhance the state-of-the-art methods for the literature-based knowledge discovery.

References

[1] Petrič, I.; Urbančič, T.; Cestnik, B. (2007). Discovering hidden knowledge from biomedical literature. *Informatica* 31(1), pp. 15-20.

[2] Petrič, I.; Urbančič, T.; Cestnik, B. (2006). Literature mining: potential for gaining hidden knowledge from biomedical articles. *Proceedings of the 9th International multi-conference Information Society IS-2006*, Ljubljana, pp. 52-55.

[3] Petrič, I.; Urbančič, T.; Cestnik, B.; Macedoni-Lukšič, M. (2009). Literature mining method RaJoLink for uncovering relations between biomedical concepts. *Journal of Biomedical Informatics* 42(2), pp. 219-227.

[4] Swanson, D. R. (1986). Undiscovered public knowledge. *Library Quarterly* 56(2), pp. 103-118.

[5] Urbančič, T.; Petrič, I.; Cestnik, B. (2009). A method for finding seeds of future discoveries in nowadays literature. *Foundations of intelligent systems.* Springer, Berlin, pp. 129-138.

[6] Urbančič, T.; Petrič, I.; Cestnik, B.; Macedoni-Lukšič, M. (2007). Literature mining: towards better understanding of autism. *Proceedings of the 11th Conference on Artificial Intelligence in Medicine in Europe*, Amsterdam, pp. 217-226.