Literature-Based Discovery by Learning Heterogeneous Bibliographic Information Networks

Yakub Sebastian
School of Information Technology
Monash University Malaysia
Malaysia
yakub.sebastian@gmail.com

Abstract

Literature-based discovery (LBD) research aims at finding effective computational methods for retrieving previously unknown information connections between clusters of research papers from disparate research areas. Existing methods encompass two general approaches. The first approach searches for these unknown connections by examining the textual contents of research papers. In addition to the existing textual features, the second approach incorporates structural features of scientific literatures, such as citation structures. These approaches, however, have not considered research papers' latent bibliographic metadata structures as important features that can be used for predicting previously unknown relationships between them.

This thesis investigates a new graph-based LBD method that exploits the latent bibliographic metadata connections between pairs of research papers. The heterogeneous bibliographic information network is proposed as an efficient graph-based data structure for modeling the complex relationships between these metadata. In contrast to previous approaches, this method seamlessly combines textual and citation information in the form of path-based metadata features for predicting future co-citation links between research papers from disparate research fields. The results reported in this thesis provide evidence that the method is effective for reconstructing the historical literature-based discovery hypotheses.

This thesis also investigates the effects of semantic modeling and topic modeling on the performance of the proposed method. For semantic modeling, a general-purpose word sense disambiguation technique is proposed to reduce the lexical ambiguity in the title and abstract of research papers. The experimental results suggest that the reduced lexical ambiguity did not necessarily lead to a better performance of the method. This thesis discusses some of the possible contributing factors to these results.

Finally, topic modeling is used for learning the latent topical relations between research papers. The learned topic model is incorporated into the heterogeneous bibliographic information network graph and allows new predictive features to be learned. The results in this thesis suggest that topic modeling improves the performance of the proposed method by

increasing the overall accuracy for predicting the future co-citation links between disparate research papers.

Supervisors: Sylvester Olubolu Orimaye, Siew Eu-Gene (Monash University Malaysia)
Available at: https://figshare.com/articles/Literature-Based_Discovery_by_Learning_Heterogeneous_Bibliographic_Information_Networks/4628671