Identifying Descriptive Keyphrases from Scholarly Big Data

Cornelia Caragea
Computer Science and Engineering
University of North Texas
Denton, TX 76203
ccaragea@unt.edu

Abstract

The large and growing amounts of online textual data present both challenges and opportunities to enhance knowledge discovery. One important challenge is to automatically extract a small set of keyphrases from a document that can accurately describe the document's content and can facilitate fast information processing. In this paper, we explore artificial intelligence approaches to keyphrase extraction from scientific research articles and study what types of information can aid keyphrase extraction and what aspects of this task seem more difficult.

1 Introduction

The current Scholarly Web contains many millions of scientific documents. For example, PubMed has over 20 million documents, whereas Google Scholar is estimated to have more than 100 millions. Open-access digital libraries such as CiteSeer^x, which acquire freely-available research articles from the Web, witness an increase in their document collections as well. These rapidly-growing scholarly document collections offer benefits for knowledge discovery, learning, and staying up-to-date with recent research advances. However, navigating in these digital libraries and finding useful information have become very challenging. Fortunately, keyphrases associated with a document typically provide a high-level topic description of the document and allow for efficient processing of more information in less time. Despite their strong value, manually annotated keyphrases are not always provided with the documents, but they need to be gleaned from the content of documents. Hence, accurate approaches for automatic keyphrase extraction from research documents are highly needed.

Previous approaches to automatic keyphrase extraction are capable of extracting keyphrases at large scale [3]. However, they are limited in terms of the types of information used. That is, this previous research has focused mainly on approaches that use only the textual content of a document or incorporate information from its textually-similar neighbors. We posit that, in addition to a document's textual content and textually-similar neighbors, other informative neighborhoods exist in document networks that have the potential to improve large-scale keyphrase extraction. For example, in a scholarly domain, research papers are not isolated. Rather, they are highly inter-connected in giant *citation networks*, in which papers *cite* one another. In a citation network, information flows from one paper to another via the citation relation. This information flow and the influence of one paper on another are specifically captured by means of *citation contexts*, i.e., short text segments surrounding a citation's mention. These contexts are not arbitrary, but they serve as brief summaries of a cited paper. Figure 1 shows an anecdotal example illustrating this behavior using the 2010 best paper award winner in the World Wide Web conference (Paper 1) and its citation network neighbor (Paper 2). Notice the large overlap in the author specified keywords and the citation contexts. To this end, more powerful methods that incorporate other types of information such as citation network information and can yet solve large-scale keyphrase extraction from research papers are highly desirable.

To address this challenge, in this paper we explore AI approaches that use information from citation contexts in novel ways to improve keyphrase extraction in both supervised and unsupervised settings.

30th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain.

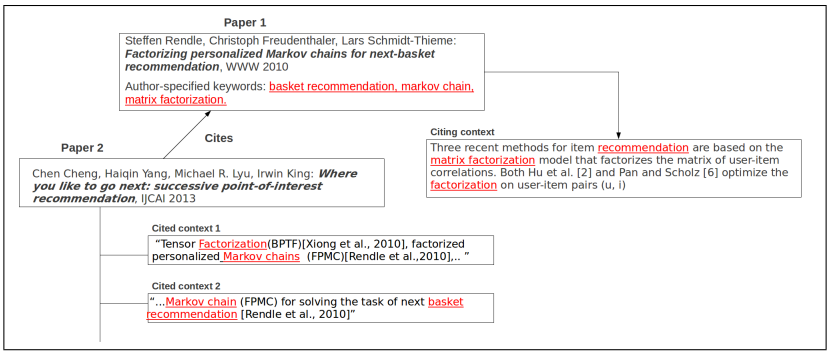

Figure 1: A small citation network.

2 Citation-based Methods for Keyphrase Extraction from Research Papers

First, we present a supervised approach to keyphrase extraction from research papers. Supervised keyphrase extraction is formulated as a binary classification problem, where candidate phrases are classified as either keyphrases or non-keyphrases. Our supervised approach effectively incorporates, in the learned models, information from the paper's local neighborhood available in citation networks. We design novel features for keyphrase extraction based on citation context information and use them in conjunction with traditional features in a supervised probabilistic framework [1]. We compare our results against the author annotated keyphrases and show empirically that the proposed models significantly outperform strong baselines on two datasets compiled from two machine learning conferences: the World Wide Web and Knowledge Discovery from Data.

Second, we design a fully unsupervised graph-based algorithm that incorporates evidence from multiple sources (citation contexts and document content) in a flexible manner [2]. Unsupervised keyphrase extraction is formulated as a ranking problem with graph-based ranking techniques being considered state-of-the-art. These techniques construct a word graph for each target document, in which nodes correspond to words and edges correspond to word association patterns. Nodes are then ranked using graph centrality measures such as PageRank and the top ranked phrases are returned as keyphrases. Unlike simple graph edges with fixed weights used in previous works, in our graph-based algorithm, we used parameterized edge weights. We showed experimentally on several datasets significant improvements over existing state-of-the-art models for keyphrase extraction. Our approach improves precision at rank 1 by as much as 9-20% over state-of-the-art baselines.

3 Discussion and Conclusion

We proposed supervised and unsupervised techniques for keyphrase extraction from research papers that are capable of incorporating complex information available in citation network. Current scholarly digital libraries such as CiteSeer^{x} make it possible to use citation network information. Through our research, we identified several aspects of this task that bring additional challenges [4]. Among these challenges, we highlight the following: the annotation process is extremely subjective (e.g., some authors use "Twitter" whereas others may prefer "social networks" to refer to the same concept); and the set of author-annotated keyphrases may be incomplete and/or noisy.

Acknowledgments

This research is supported by NSF award #1423337 to Cornelia Caragea. Any opinions, findings, and conclusions expressed here are those of the authors and do not necessarily reflect the views of NSF.

References

[1] Cornelia Caragea, Florin Bulgarov, Andreea Godea, and Sujatha Das Gollapalli. Citation-enhanced keyphrase extraction from research papers: A supervised approach. In *Proceedings of EMNLP*, 2014.

[2] Sujatha Das Gollapalli and Cornelia Caragea. Extracting keyphrases from research papers using citation networks. In *Proceedings of AAAI*, pages 1629–1635, 2014.

[3] Kazi Saidul Hasan and Vincent Ng. Automatic keyphrase extraction: A survey of the state of the art. In *Proceedings of ACL*, 2014.

[4] Lucas Sterckx, Cornelia Caragea, Thomas Demeester, and Chris Develder. Supervised keyphrase extraction as positive unlabeled learning. In *Proceedings of EMNLP*, 2016.