Keyphrase Extraction for Scholarly Big Data

Cornelia Caragea

Computer Science and Engineering
University of North Texas

July 10, 2015

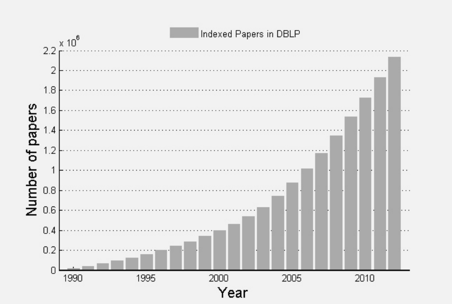
MLg Machine Learning Group
UNT Computer Science and Engineering

Scholarly Big Data

Large number of scholarly documents on the Web

- PubMed currently has over 24 million documents
- Google Scholar is estimated to have 160 million documents

The growth in the number of papers indexed DBLP:



- Navigating in these digital libraries has become very challenging.

Keyphrase Extraction in Document Networks

- Keyphrases:
  - Allow for efficient processing of more information in less time
  - Are useful in many applications:
    - Topic tracking, information filtering and search, classification, clustering, and recommendation.
- Keyphrase extraction is the task of automatically extracting descriptive phrases or "concepts" from a document.

Keyphrases

Example: A snippet from the 2010 best paper award winner in the WWW conference - the author-input keyphrases are shown in red

Factorizing Personalized Markov Chains for Next-Basket Recommendation

by Rendle, Freudenthaler, and Schmidt-Thieme

"Recommender systems are an important component of many websites. Two of the most popular approaches are based on matrix factorization (MF) and Markov chains (MC). MF methods learn the general taste of a user by factorizing the matrix over observed user-item preferences. *[…]* In this paper, we present a method bringing both approaches together. Our method is based on personalized transition graphs over underlying Markov chains. *[…]* We show that our factorized personalized MC (FPMC) model subsumes both a common Markov chain and the normal matrix factorization model. For learning the model parameters, we introduce an adaption of the Bayesian Personalized Ranking (BPR) framework for sequential basket data. *[…]*"

Previous Approaches to Keyphrase Extraction

- Use generally only the textual content of the target document [Mihalcea and Tarau, 2004], [Liu et al., 2010].
- Recently, models are proposed that incorporate a local neighborhood of a document [[Wan and Xiao, 2008].
  - Obtained improvements over models that use only textual content.
  - However, their neighborhood is limited to textually-similar documents.

During these "Big Data" times - access to giant document networks

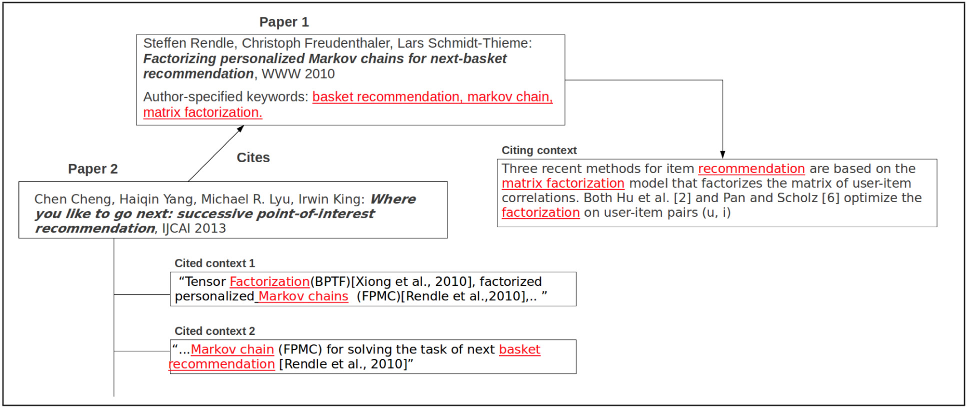
Our Questions

- In addition to a document's textual content and textually-similar neighbors, are there other informative neighborhoods in research document collections?
- Can these neighborhoods improve keyphrase extraction?

From Data to Knowledge

- A typical scientific research paper:
  - Proposes new problems or extends the state-of-the-art for existing research problems.
  - Cites relevant, previously-published papers in appropriate *contexts.*
- Citation contexts capture the influence of one paper on another as well as the flow of information
  - Can serve as "micro summaries" of a cited paper!

A Small Citation Network



- Citation contexts are very informative!

[Das G. and Caragea, 2014]; [Caragea et al., 2014]

Citation Contexts - Not a New Idea

- Using terms from citation contexts resembles the analysis of hyperlinks and the graph structure of the Web
  - Web search engines build on the intuition that the anchor text pointing to a page is a good descriptor of its content, and thus use anchor terms as additional index terms for a target webpage.
- Previously used for other tasks:
  - Scientific paper summarization [Mei and Zhai, 2008; Abu-Jbara and Radev, 2011; Qazvinian et al., 2010]
  - Indexing of cited papers [Ritchie et al. (2006)]
  - Author influence in document networks [Kataria et al., 2011]

Citation Contexts to Keyphrase Extraction

- How can we use these contexts and how do they help in keyphrase extraction?
- We proposed:
  - **CiteTextRank** [Das Gollapalli and Caragea, 2014]: an unsupervised, graph-based algorithm that incorporates evidence from multiple sources (citation contexts as well as document content) in a flexible way to extract keyphrases.
  - **Citation-enhanced Keyphrase Extraction (CeKE)** [Caragea et al., 2014]: a supervised binary classification model built on a combination of novel features that capture information from citation contexts and existing features from previous works.

Features for CeKE

Feature Name	Description
Existing features for keyphrase extraction	
tf-idf	term frequency * inverse document frequency, computed from a target paper; used in KEA
relativePos	the position of first occurrence of a phrase divided by the total number of tokens; used in KEA and Hulth's methods
POS	the part-of-speech tag of the phrase; used in Hulth's methods
Novel features - Citation Network Based	
inCited	if the phrase occurs in cited contexts
inCiting	if the phrase occurs in citing contexts
citation tf-idf	the *tf-idf* value of the phrase, computed from the aggregated citation contexts
Novel features - Extensions of Existing Features	
first position	the distance of the first occurrence of a phrase from the beginning of a paper
tf-idf-Over	*tf-idf* larger than a threshold θ
firstPosUnder	the distance of the first occurrence of a phrase from the beginning of a paper is below some value β

How Does CeKE Compare with Supervised Models?

	WWW			KDD		
Method	Precision	Recall	F1-score	Precision	Recall	F1-score
Citation - Enhanced (CeKE)	**0.227**	**0.386**	**0.284**	**0.213**	**0.413**	**0.280**
Hulth - n-gram with tags	0.165	0.107	0.129	0.206	0.151	0.172
KEA	0.210	0.146	0.168	0.178	0.124	0.145

Table: Comparison of CeKE with Hulth's and KEA methods.

Features used in previous supervised methods:

- Hulth's features: *POS*, *relative position*, *term frequency* and *collection frequency*.
- KEA's features: *tf-idf* and *relative position*

How Does CeKE Perform in the Absence of Either Cited or Citing Contexts?

	WWW			KDD		
Method	Precision	Recall	F1-score	Precision	Recall	F1-score
CeKE - Both contexts	**0.227**	**0.386**	**0.284**	**0.213**	**0.413**	**0.280**
CeKE - Only cited contexts	0.222	0.286	0.247	0.192	0.300	0.233
CeKE - Only citing contexts	0.203	0.342	0.253	0.195	0.351	0.250

Table: Results with both contexts and only cited/citing contexts.

Conclusions and Future Directions

- Our models give significant improvements over baseline models for multiple datasets of research papers in the Computer Science domain
- Future directions:
    - Citation context lengths: Incorporate more sophisticated approaches to identifying the text that is relevant to a target citation [Abu-Jbara and Radev, 2012; Teufel, 1999] and study the influence of context lengths on the quality of extracted keyphrase
    - Evaluate our models on other domains, e.g., the ACL Anthology, PubMed.

References

- S. Das Gollapalli and C. Caragea (2014). Extracting Keyphrases from Research Papers using Citation Networks. In: *Proceedings of the 28th American Association for Artificial Intelligence (AAAI '14).*
- C. Caragea, F. Bulgarov, A. Godea, and S. Das Gollapalli. (2014). Citation-Enhanced Keyphrase Extraction from Research Papers: A Supervised Approach. In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP '14).*
- R. Mihalcea and P. Tarau. (2004). TextRank: Bringing order into text. In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP '04).*
- X. Wan and J. Xiao (2008). Single document keyphrase extraction using neighborhood knowledge. In: *Proceedings of the Association for the Advancement of Artificial Intelligence (AAAI '08).*
- Qiaozhu Mei and ChengXiang Zhai. (2008). Generating impact-based summaries for scientific literature. In: *Proceedings of the Conference of the Association for Computational Linguistics, pages 816-824, Columbus, Ohio.*
- V. Qazvinian, D. Radev, and A. Özgür. 2010. Citation summarization through keyphrase extraction. In: *Proceedings of the 23rd International Conference on Computational Linguistics, COLING '10, pages 895 - 903.*

ACL Workshop on Keyphrase Extraction

Novel Computational Approaches to Keyphrase Extraction
Workshop co-located with ACL 2015

- For more information, please visit:
  www.cse.unt.edu/∼ccaragea/acl2015ws.html

Thank you!

NSF National Science Foundation
WHERE DISCOVERIES BEGIN

CiteSeerX =5M