Keyphrase Extraction from Scholarly Documents for Data Discovery and Reuse

Cornelia Caragea and C. Lee Giles

Artificial Intelligence for Data Discovery and Reuse

May 13, 2019

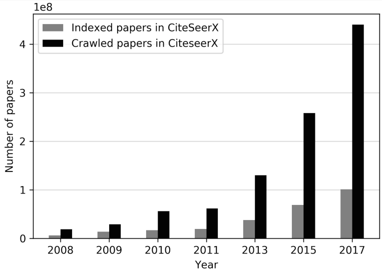
Information Retrieval Group
UIC Computer Science

Scholarly Big Data

Large number of scholarly documents on the Web

- Microsoft Academic expanded from 83 million records in 2015 to 168 million in 2017 [Hug and Brandle, 2017].
- Google Scholar was estimated to have ≈ 160 million documents in 2014 [Orduna-Malea et al, 2015].

The growth in the number of papers crawled and indexed by CiteSeerX:



- Navigating in these digital libraries has become very challenging.

Keyphrases

- **Keyphrases** provide a high-level topic description of a document and can allow for *efficient* processing of more information in less time

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 Steffen Rendle, Christoph Freudenthaler and Lars 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. [...]

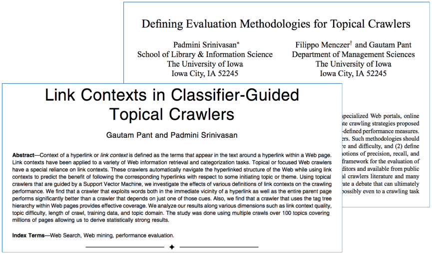
The Reuse of Keyphrases

- Keyphrases associated with research papers are **reused** in many other applications...

Characterizing an Author

Matching Reviewers with Submissions

Performance-Evaluation
Web-Mining
Search
Topical-Crawlers
Web-Search



Defining Evaluation Methodologies for Topical Crawlers

Padmini Srinivasan*
School of Library & Information Science
The University of Iowa
Iowa City, IA 52245

Filippo Menczer† and Gautam Pant
Department of Management Sciences
The University of Iowa
Iowa City, IA 52245

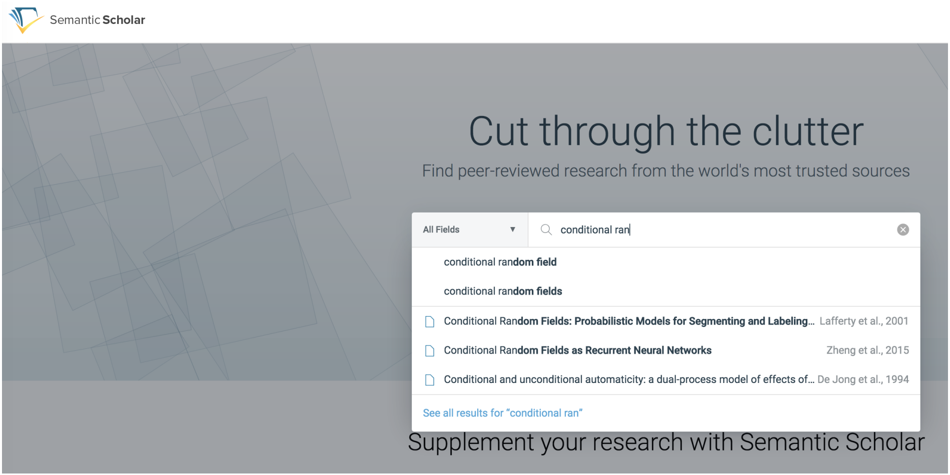
Link Contexts in Classifier-Guided Topical Crawlers

Gautam Pant and Padmini Srinivasan

Abstract—Context of a hyperlink or *link context* is defined as the terms that appear in the text around a hyperlink within a Web page. Link contexts have been applied to a variety of Web information retrieval and categorization tasks. Topical or focused Web crawlers have a special reliance on link contexts. These crawlers automatically navigate the hyperlinked structure of the Web while using link contexts to predict the benefit of following the corresponding hyperlinks with respect to some initiating topic or theme. Using topical crawlers that are guided by a Support Vector Machine, we investigate the effects of various definitions of link contexts on the crawling performance. We find that a crawler that exploits words both in the immediate vicinity of a hyperlink as well as the entire parent page performs significantly better than a crawler that depends on just one of those cues. Also, we find that a crawler that uses the tag tree hierarchy within Web pages provides effective coverage. We analyze our results along various dimensions such as link context quality, topic difficulty, length of crawl, training data, and topic domain. The study was done using multiple crawls over 100 topics covering millions of pages allowing us to derive statistically strong results.

Index Terms—Web Search, Web mining, performance evaluation.

Keyword Suggestion for Query Formulation



The Reuse of Keyphrases

- Keyphrases associated with research papers are also useful in applications such as:
  - Topic tracking
  - Information filtering and search
  - Query expansion
  - Document clustering, classification, and summarization
  - Reading comprehension...

Keyphrases for Data Discovery

- Keyphrases are also useful for data discovery in digital library applications.

Document Discovery

topic modeling with network regularization

All News Videos Images Shopping More Settings Tools

About 1,380,000 results (0.33 seconds)

Scholarly articles for **topic modeling with network regularization**

Topic modeling with **network regularization** - Mei - Cited by 415

Relational **topic models** for document **networks** - Chang - Cited by 457

Probabilistic **topic models** with biased propagation on ... - Deng - Cited by 115

[PDF] Topic Modeling with Network Regularization - umich.edu and www ...

www-personal.umich.edu/~qmei/pub/www08-netplsa.pdf ▾

by Q Mei - Cited by 414 - Related articles

The proposed method combines **topic** mod- eling and social **network** analysis, and leverages the power of both statistical **topic models** and discrete **regularization**. ... The proposed **model** is general; it can be applied to any text collections with a mixture of topics and an associ- ated **network** structure.

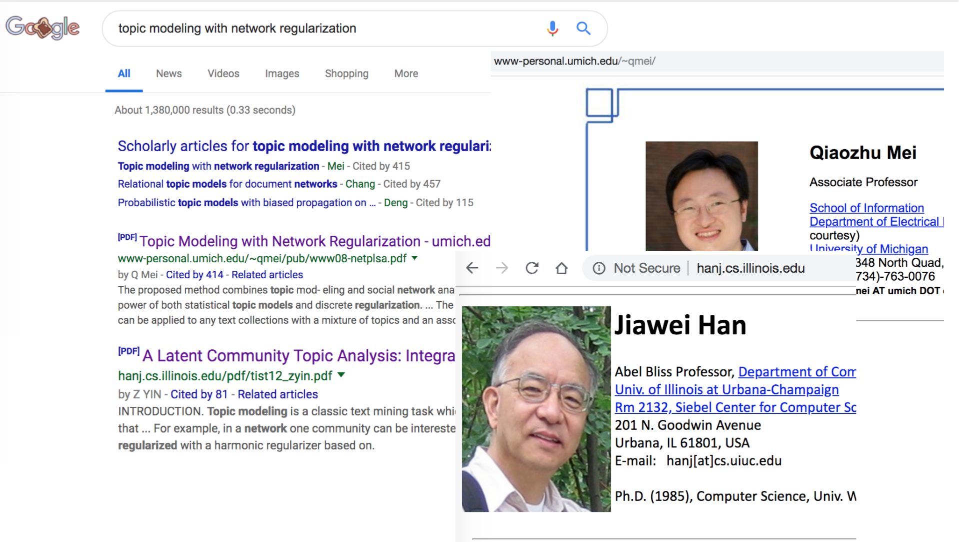
[PDF] A Latent Community Topic Analysis: Integration of ... - Jiawei Han

hanj.cs.illinois.edu/pdf/tist12_zyin.pdf ▾

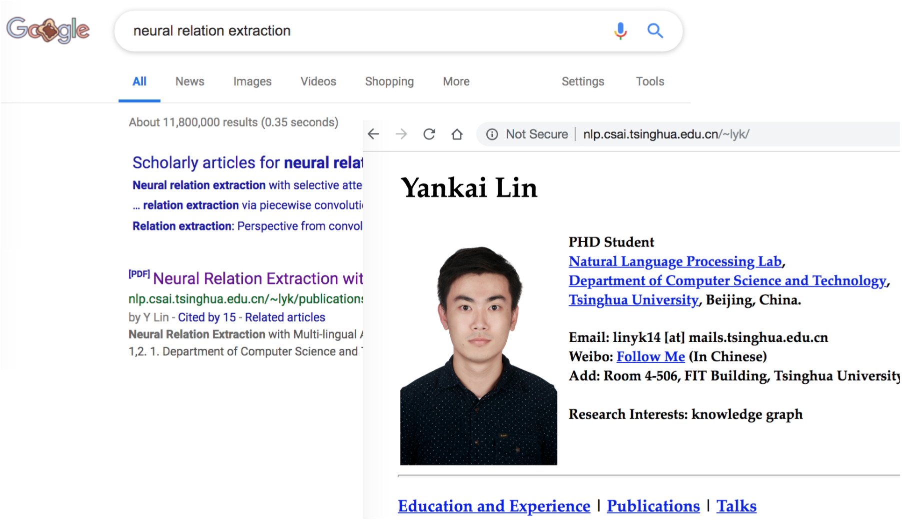
by Z YIN - Cited by 81 - Related articles

INTRODUCTION. **Topic modeling** is a classic text mining task which is to discover the hidden topics that ... For example, in a **network** one community can be interested in both ...... 2008]: PLSA **regularized** with a harmonic regularizer based on.

Author Homepage Discovery



Author Homepage Discovery



Keyphrase Extraction

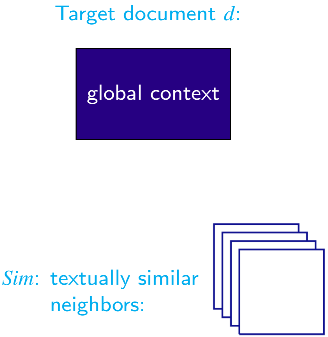
- Despite their importance, manually annotated keyphrases are not always provided with the documents:
  - Need to be gleaned from the content of documents.
  - E.g., documents available from the ACL Anthology.
- Hence, accurate approaches are required for **keyphrase extraction** from research documents
  - **Keyphrase extraction** is defined as the problem of automatically extracting descriptive phrases or concepts from documents

Previous Approaches to Keyphrase Extraction

- Many approaches have been studied [Hasan and Ng, 2014]:
  - Supervised approaches [Medelyan et al., 2009; Hulth, 2003; Turney, 2000]
    - Binary classification: candidate phrases classified as keyphrases or non-keyphrases.
  - Unsupervised approaches [Florescu and Caragea, 2017; Liu et al., 2010; Wan and Xiao, 2008; Mihalcea and Tarau, 2004]
    - Ranking: candidate phrases are ranked using various measures such as tf, tf-idf, and PageRank scores.
  - Neural approaches [Al-Zaidy, Caragea, Giles, 2019; Gollapalli et al, 2018; Meng et al., 2017]
    - Sequence to sequence models or sequence labeling with a Conditional Random Fields layer.

Limitations of Previous Approaches

- Generally, previous approaches:
  - Use only the textual content of the target document [Mihalcea and Tarau, 2004; Liu et al., 2010].
  - Incorporate a local neighborhood of a document for extracting keyphrases [Wan and Xiao, 2008]
    - However, the neighborhood is limited to textually-similar documents.



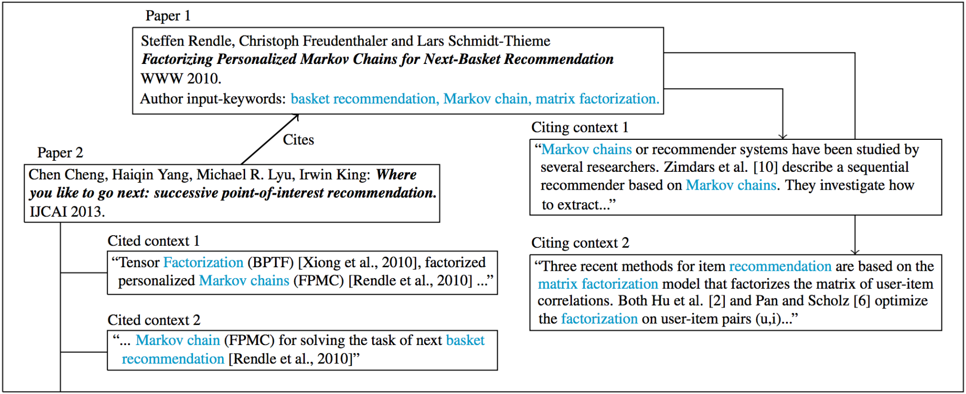
Our Questions

- **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 in large citation networks and serve as "micro summaries" of a cited paper!

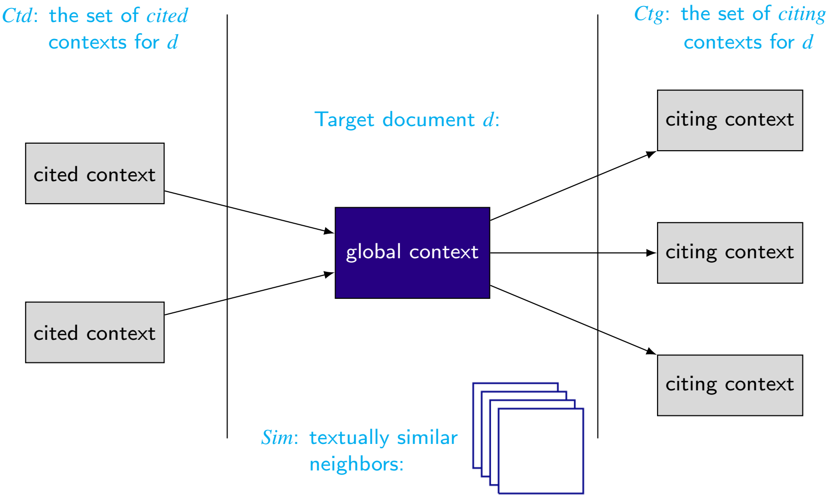
A Small Citation Network



- Citation contexts are very informative!

[Gollapalli and Caragea, 2014 (**AAAI**); Caragea, 2016 (**AI4DataSci**)]

Citation Contexts for Keyphrase Extraction



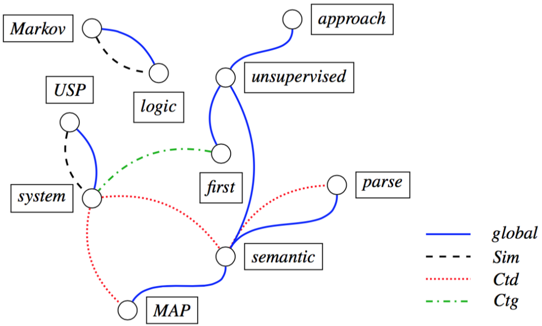
- $T = \{Ctd, Ctg, Sim, g\}$ represents the types of available contexts for d.

CiteTextRank: An Unsupervised Approach

Unsupervised Semantic Parsing

We present the first unsupervised approach to the problem of learning a semantic parser, using Markov logic . Our USP system transforms dependency trees into quasi-logical forms, recursively induces lambda forms from these, and clusters them to abstract away syntactic variations of the same meaning. The MAP semantic parse of a sentence is obtained by recursively assigning its parts to lambda-form clusters and composing them. We evaluate our approach by using it to extract a knowledge base from biomedical abstracts and answer questions. USP substantially outperforms TextRunner, DIRT and an informed baseline on both precision and recall on this task.

$w = 2$:



[Gollapalli and Caragea, 2014 (**AAAI**)]

CeKE: A Supervised Approach

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

[Caragea et al., 2014 (**EMNLP**); Bulgarov and Caragea, 2015 (**WWW**)]

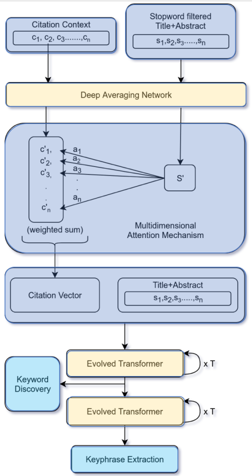
Supervised vs. Unsupervised Models

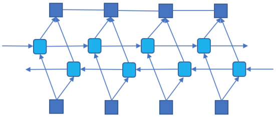
	WWW			KDD		
Method	Precision	Recall	F1-score	Precision	Recall	F1-score
Supervised						
Citation - Enhanced (CeKE)	**0.227**	**0.386**	**0.284**	**0.213**	**0.413**	**0.280**
Hulth - n-gram with tags	0.165	0.107	0.129	0.206	0.151	0.172
KEA	0.210	0.146	0.168	0.178	0.124	0.145
Unsupervised - Top 5 predicted keyphrases						
CiteTextRank	**0.110**	**0.134**	**0.119**	**0.133**	**0.153**	**0.141**
TF-IDF	0.089	0.100	0.094	0.083	0.102	0.092
TextRank	0.058	0.071	0.062	0.051	0.065	0.056
ExpandRank - 1 neigh.	0.088	0.109	0.095	0.077	0.103	0.086
ExpandRank - 5 neigh.	0.093	0.113	0.100	0.080	0.108	0.090

[Gollapalli and Caragea, 2015 (**AAAI**); Caragea et al., 2014 (**EMNLP**)]

Neural Models

- Universal Evolved Transformer in a Multi-Task Learning Framework with Integration of Information from Citation Contexts.





Bi-LSTM-CRF

[Al-Zaidy, Caragea, and Giles, 2019 (**WWW**)]

Model	**Pr**	**Re**	$\mathbf{F_1}$
ACM Data			
Bi-LSTM-CRF	**34.42%**	36.07%	35.22%
Bi-LSTM-MTL	28.4%	46.96%	35.44%
UT-MTL	29.3%	42.16%	34.6%
ET-MTL	30.93%	**46.63%**	37.19%
ET-MTL + CITATIONS	33.72%	44.73%	**38.45%**

[Ray Chowdhury and Caragea, 2019 (Submitted)]

Summary

- Developments in keyphrase extraction are central to ***knowledge discovery and organization*** and have a direct impact on the development of digital libraries.
- Our major contribution was to integrate citation contexts for keyphrase extraction.
  - ***Our model outperforms strong baselines in terms of all performance measures on scholarly documents***
- Future directions:
  - Extend our models to other CS areas and other scientific domains, e.g., PubMed, Social Science, Political Science, Ecology.
  - Predict terms not found in a target paper to be keyphrases (through semantic and syntactic features).

References

- R. Al-Zaidy, C. Caragea, L. Giles (2019). Bi-LSTM-CRF Sequence Labeling for Keyphrase Extraction from Scholarly Documents.. In: *Proc. of The Web Conference (**WWW '19**).*

- C. Florescu and C. Caragea (2017). PositionRank: An Unsupervised Approach to Keyphrase Extraction from Scholarly Documents. In: *Proc. of the Annual Meeting of the Association for Computational Linguistics (**ACL '17**).*

- C. Caragea (2016). Identifying Descriptive Keyphrases from Scholarly Big Data. In: *Artificial Intelligence for Data Science (**AI4DataSci '16**).*

- L. Sterckx, C. Caragea, T. Demeester, and C. Develder. (2016). Supervised Keyphrase Extraction as Positive Unlabeled Learning. In: *Proc. of the Conference on Empirical Methods in Natural Language Processing (**EMNLP '16**).*

- S. Das Gollapalli and C. Caragea (2014). Extracting Keyphrases from Research Papers using Citation Networks. In: *Proc. 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: *Proc. of the Conference on Empirical Methods in Natural Language Processing (**EMNLP '14**).*

Thank you!

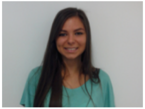
- **Acknowledgements:**

National Science Foundation
WHERE DISCOVERIES BEGIN

CiteSeer^{x}


Florin Bulgarov


Sujatha Das


Andreea Godea


Jishnu Chowdhury


Krutarth Patel


Lucas Sterckx


Corina Florescu


Alina Ciobanu


Ana Uban


Kishore Neppalli

IRg Information Retrieval Group
UIC Computer Science