Citation-Enhanced Keyphrase Extraction from Research Papers: A Supervised Approach

Cornelia Caragea, Florin Bulgarov, Andreea Godea and Sujatha Das Gollapalli

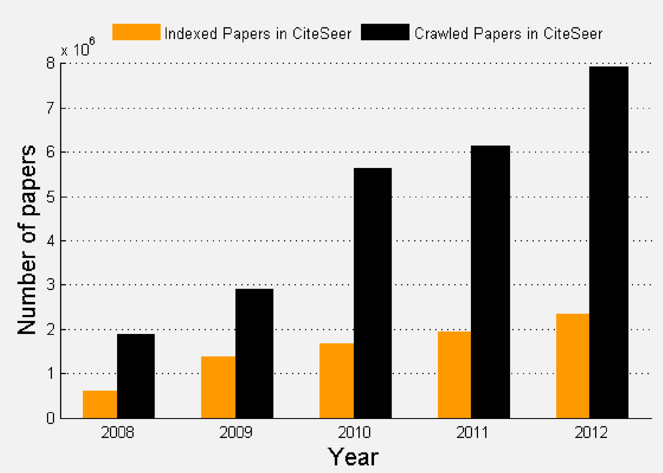
Computer Science and Engineering, University of North Texas

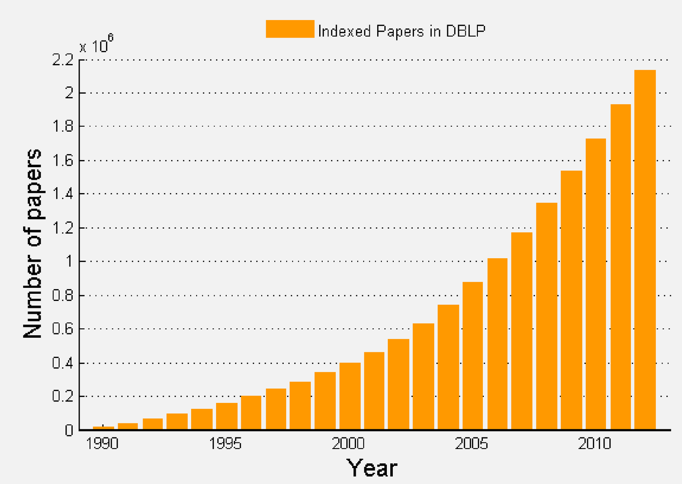
Why Keyphrase Extraction?

- The number of scholarly documents on the Web is exponentially increasing every year.
- Keyphrase extraction is the problem of automatically extracting important phrases or concepts (i.e. the *essence*) of a document.
- Keyphrases are shown to be rich sources of information for many applications such as document classification, clustering, recommendation, indexing, searching and summarization.
- During these *big data* times, keyphrases associated with research papers can allow for *efficient processing of more information in less time*.
  - Also useful in many ML and IR applications such as topic tracking, information filtering, and search.

Stats

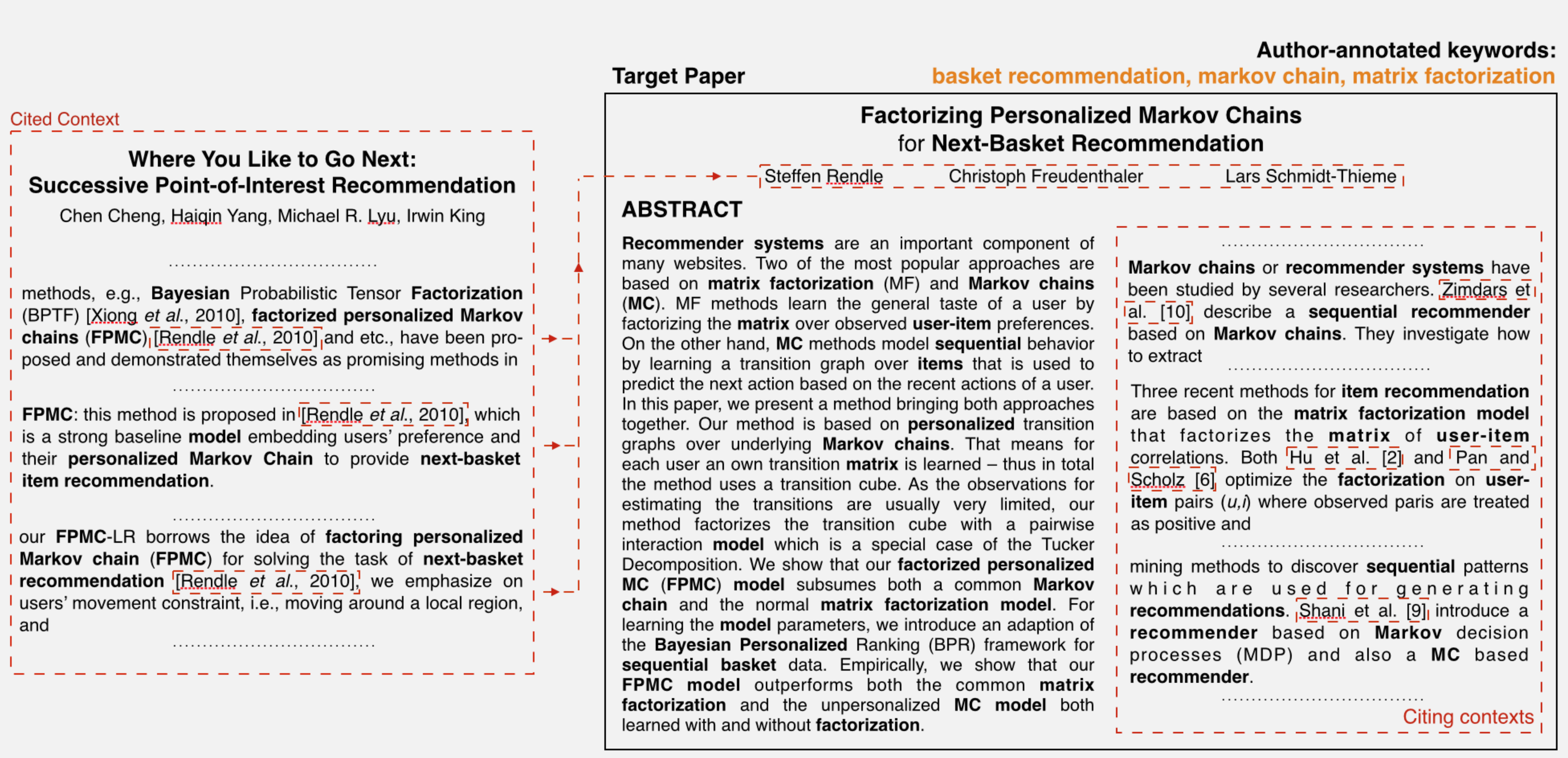
- Number of Research Papers indexed by two important digital libraries in the fields of computer and information science over the past years.





From Data to Knowledge

- Scientific research papers typically propose innovative solutions or extend the state-of-the-art algorithms for existing research problems.
- In addition to a document's textual content, other informative neighborhoods exist that have the potential to improve keyphrase extraction. For example, research papers are highly inter-connected in giant *citation networks*, where papers *cite* or *are cited* by others.
  - In a citation network, information flows from one paper to another via the citation relation.
  - These contexts are not arbitrary, but they serve as brief summaries of a cited paper.



Can we effectively exploit information available in large inter-linked document networks in order to improve the performance of keyphrase extraction?

Citation-Enhanced Keyphrase Extraction (CeKE)

We propose a supervised binary classification model called CeKE, built on a combination of novel features that capture information from citation contexts and existing features from previous works.

- Generating Candidate Phrases
  - We first apply parts-of-speech filters and retain only the nouns and adjectives
  - Porter Stemmer is applied on every word
  - Words that have contiguous positions in the document are concatenated into n-grams
  - Finally, we eliminate phrases that end with an adjective and the unigrams that are adjectives
- Features

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

Datasets

- We compiled two datasets consisting of titles and abstracts from two top-tier machine learning conferences:
  - World Wide Web (WWW)
  - Knowledge Discovery and Data Mining (KDD)
- The *author-annotated* keyphrases were treated as the gold standard.
- The citation contexts' length was set to 50 words around the citation mention.

Dataset	Num. (#) Papers	Average Cited Ctx.	Average Citing Ctx.	Average Keyphrases	#uni-grams	#bi-grams	#tri-grams
WWW	425	15.45	18.78	4.87	680	1036	247
KDD	365	12.69	19.74	4.03	363	853	189

Results

How does CeKE compare with the existing supervised models that use only information intrinsic to the data?

	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

Figure : Comparison of CeKE with Hulth's n-grams with tags and KEA methods.
Hulth's features: *POS, relative position, document frequency* and *collection frequency*.
KEA's features: *tf-idf* and *relative position*

Results

How is our Citation-Enhanced algorithm comparing with recent unsupervised 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**
TF-IDF - Top 5	0.089	0.100	0.094	0.083	0.102	0.092
TF-IDF - Top 10	0.075	0.169	0.104	0.080	0.203	0.115
TextRank - Top 5	0.058	0.071	0.062	0.051	0.065	0.056
TextRank - Top 10	0.062	0.133	0.081	0.053	0.127	0.072
ExpandRank - 1 neigh. - Top 5	0.088	0.109	0.095	0.077	0.103	0.086
ExpandRank - 1 neigh. - Top 10	0.078	0.165	0.101	0.071	0.177	0.098
ExpandRank - 5 neigh. - Top 5	0.093	0.113	0.100	0.080	0.108	0.090
ExpandRank - 5 neigh. - Top 10	0.080	0.172	0.104	0.068	0.172	0.095
ExpandRank - 10 neigh. - Top 5	0.094	0.113	0.100	0.077	0.103	0.086
ExpandRank - 10 neigh. - Top 10	0.076	0.162	0.099	0.065	0.164	0.091

Figure : Comparison of CeKE state-of-the-art unsupervised systems.
TextRank: window size is set to 2. *ExpandRank*: window size is set to 10.

How well does our proposed model 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

Figure : Results of CeKE using both context and using only cited or citing contexts

Anecdotal Evidence

- Our classifier was trained on both WWW and KDD datasets and was evaluated on an award winning paper published in the EMNLP Conference.
- We gathered from the Web 49 cited contexts and 30 citing contexts.
- The classifier was tuned to return as keyphrases only those that had an extremely high probability (we used a threshold of 0.985).

Unsupervised Semantic Parsing$^{0.997}$

We present the first unsupervised approach to the problem of learning a **semantic parser**$^{1.000}$, using **Markov logic**$^{0.991}$. Our **USP system**$^{0.985}$ 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**$^{1.000}$ 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**$^{1.000}$ substantially outperforms TextRunner, DIRT and an informed baseline on both precision and recall on this task.

Human annotated labels: *unsupervised semantic parsing, Markov logic, USP system*

Figure : The title and abstract of an award winning paper published in the EMNLP conference by Poon and Domingos (2009). *Grey* - filtered out words; *Black* - candidate phrases; ***Bold red*** - predicted keyphrases; *Numbers* - classifier's confidence.

Keyphrase	#cited c.	#citing c.
semantic parser	29	26
USP	31	10
Markov logic	15	10
unsupervised semantic parsing	12	1
USP system	3	2

The table on the left shows the term-frequency of every predicted keyphrase within the citation network.

Conclusions and Future Directions

- The proposed citation-enhanced supervised framework performs substantially better compared with state-of-the-art supervised and unsupervised models.
- Our model can be extended to other types of documents such as webpages, weblogs or e-mails.
- Using external sources, e.g. Wikipedia, could increase the performance by identifying better candidate phrases.
- Extensions to other domains, e.g. Biology and Chemistry, and other applications, e.g. document summarization, are of particular interest.

Acknowledgements: This research is partially supported by a grant from NSF to Cornelia Caragea. We thank Kishore Neppalli for helping with dataset construction tasks.

WWW: http://www.cse.unt.edu/~ccaragea/keyphrases.html