Keyphrase Extraction for Scholarly Big Data

Cornelia Caragea

Computer Science and Engineering
University of North Texas

July 10, 2015
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.
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?
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!
Citation contexts are very informative!

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

<table>
<thead>
<tr>
<th>Feature Name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Existing features for keyphrase extraction</strong></td>
<td></td>
</tr>
<tr>
<td><em>tf-idf</em></td>
<td>term frequency * inverse document frequency, computed from a target paper; used in KEA</td>
</tr>
<tr>
<td><em>relativePos</em></td>
<td>the position of first occurrence of a phrase divided by the total number of tokens; used in KEA and Hulth’s methods</td>
</tr>
<tr>
<td>POS</td>
<td>the part-of-speech tag of the phrase; used in Hulth’s methods</td>
</tr>
<tr>
<td><strong>Novel features - Citation Network Based</strong></td>
<td></td>
</tr>
<tr>
<td><em>inCited</em></td>
<td>if the phrase occurs in cited contexts</td>
</tr>
<tr>
<td><em>inCiting</em></td>
<td>if the phrase occurs in citing contexts</td>
</tr>
<tr>
<td><em>citation tf-idf</em></td>
<td>the <em>tf-idf</em> value of the phrase, computed from the aggregated citation contexts</td>
</tr>
<tr>
<td><strong>Novel features - Extensions of Existing Features</strong></td>
<td></td>
</tr>
<tr>
<td><em>first position</em></td>
<td>the distance of the first occurrence of a phrase from the beginning of a paper</td>
</tr>
<tr>
<td><em>tf-idf-Over</em></td>
<td><em>tf-idf</em> larger than a threshold $\theta$</td>
</tr>
<tr>
<td><em>firstPosUnder</em></td>
<td>the distance of the first occurrence of a phrase from the beginning of a paper is below some value $\beta$</td>
</tr>
</tbody>
</table>
How Does CeKE Compare with Supervised Models?

<table>
<thead>
<tr>
<th>Method</th>
<th>WWW Precision</th>
<th>WWW Recall</th>
<th>WWW F1-score</th>
<th>KDD Precision</th>
<th>KDD Recall</th>
<th>KDD F1-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Citation - Enhanced (CeKE)</td>
<td>0.227</td>
<td>0.386</td>
<td>0.284</td>
<td>0.213</td>
<td>0.413</td>
<td>0.280</td>
</tr>
<tr>
<td>Hulth - n-gram with tags</td>
<td>0.165</td>
<td>0.107</td>
<td>0.129</td>
<td>0.206</td>
<td>0.151</td>
<td>0.172</td>
</tr>
<tr>
<td>KEA</td>
<td>0.210</td>
<td>0.146</td>
<td>0.168</td>
<td>0.178</td>
<td>0.124</td>
<td>0.145</td>
</tr>
</tbody>
</table>

**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?

<table>
<thead>
<tr>
<th>Method</th>
<th>WWW</th>
<th></th>
<th>KDD</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Precision</td>
<td>Recall</td>
<td>F1-score</td>
<td>Precision</td>
</tr>
<tr>
<td>CeKE - Both contexts</td>
<td>0.227</td>
<td>0.386</td>
<td>0.284</td>
<td>0.213</td>
</tr>
<tr>
<td>CeKE - Only cited contexts</td>
<td>0.222</td>
<td>0.286</td>
<td>0.247</td>
<td>0.192</td>
</tr>
<tr>
<td>CeKE - Only citing contexts</td>
<td>0.203</td>
<td>0.342</td>
<td>0.253</td>
<td>0.195</td>
</tr>
</tbody>
</table>

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)*.


- **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

For more information, please visit:
www.cse.unt.edu/~ccaragea/acl2015ws.html
Thank you!