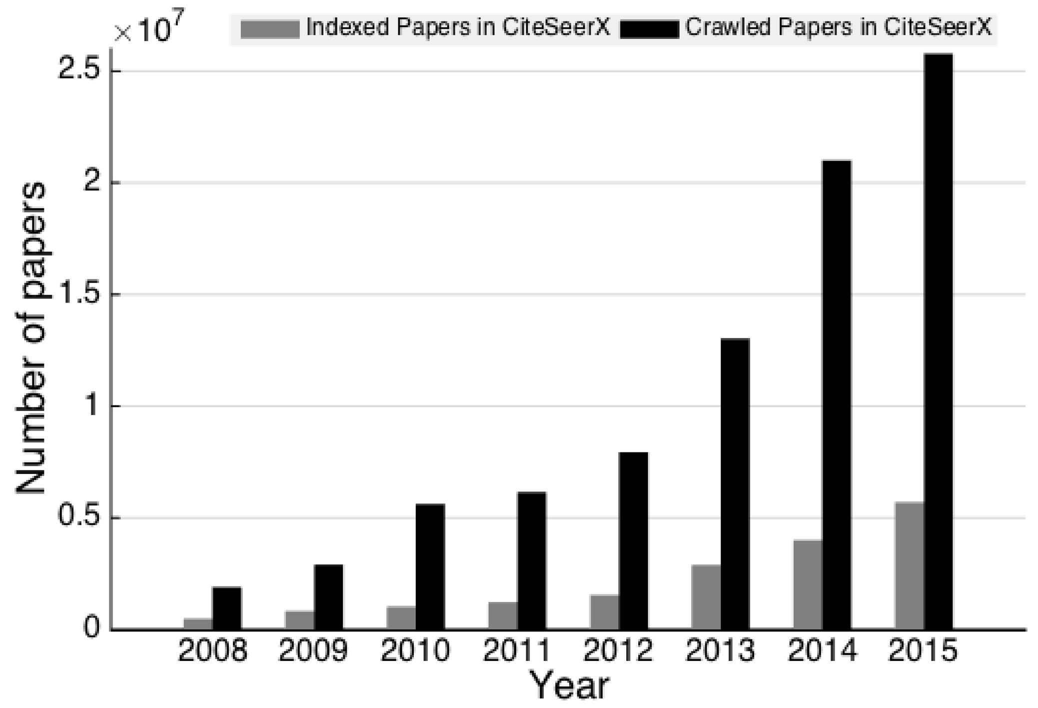
Keyphrase Extraction: An Unsupervised Approach

Corina Florescu and Cornelia Caragea

Computer Science and Engineering, University of North Texas

Why Keyphrase Extraction?

- Large and growing amounts of research articles indexed by digital libraries.



- Navigating in these digital libraries has become very challenging.
- Keyphrases of a document can allow for *efficient processing of more information in less time* and can improve many natural language processing and information retrieval tasks, e.g., summarization and contextual advertisement.
- Keyphrase extraction is defined as the problem of automatically extracting descriptive phrases or concepts from a document.

Previous Approaches to Keyphrase Extraction

- Many approaches to keyphrase extraction have been proposed in the literature along two lines of research: supervised and unsupervised.
- In the supervised line of research, different feature sets (e.g., *term frequency, relative position of the first occurrence, part-of-speech tag*) and classification algorithms (e.g., Naive Bayes) give rise to various supervised keyphrase extraction models [Hulth(2003), Caragea et al.(2014)Caragea, Bulgarov, Godea, & Gollapalli].
- Many features used to encode a candidate phrase in the supervised approaches have influenced the progress of unsupervised line of research.
  - Candidate words to be added in the graph are words with certain part of speech tags [Mihalcea & Tarau(2004), Gollapalli & Caragea(2014)] and *tf* or *tf-idf* are used to rank candidate phrases in a document [Barker & Cornacchia(2000)].
- We posit that other information can be leveraged that has the potential to improve the keyprase extraction task.

From Data to Knowledge

- Intuitively, keyphrases occur very early in a document and appear frequently.

> 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. [...]

Author-input keyphrases: *Basket Recommendation, Markov Chain, Matrix Factorization*

Figure: The title, abstract and author-input keyphrases (marked in red in the text) for the 2010 best paper award winner in the World Wide Web conference.

- *Can position information improve the performance of unsupervised keyphrase extraction task? How can we design an efficient and effective unsupervised approach for keyphrase extraction by exploiting the position information of a phrase in a document?*

Proposed Approach

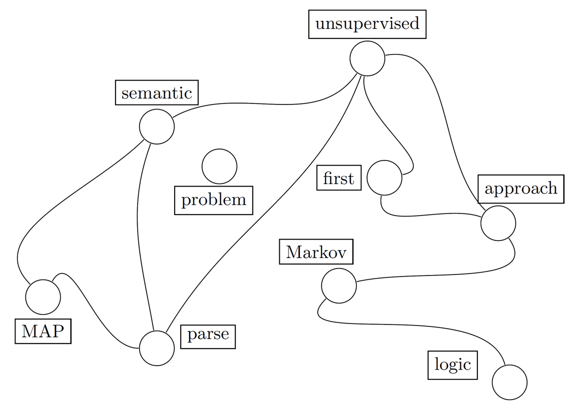
We propose a fully unsupervised graph-based algorithm that incorporates information from all positions of a word's occurrences into a biased-PageRank to score keywords that are later used to score keyphrases.

- Our approach involves there essential steps:
  1. The graph construction at the word level;
  2. The design of biased-PageRank algorithm;
  3. The scoring of multi-word phrases.

Graph Construction

> 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**.

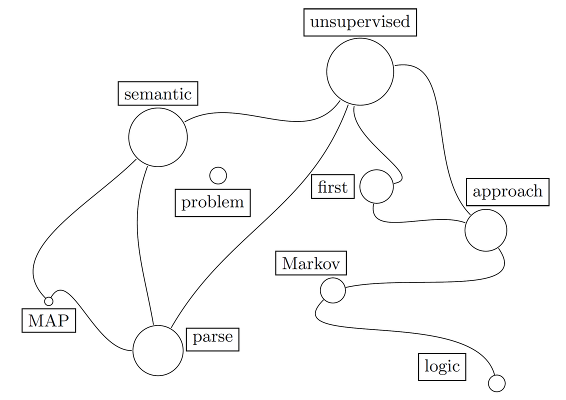
- window = 3



Position Biased PageRank

- The idea of our approach is to assign higher probabilities to those words that occur very early in the document.
- We weight each candidate word with its inverse position in the document. If the same word appears multiple times in target document, then we add all its position weights.
- Similar to Haveliwala [Haveliwala(2002)], we biased PageRank to prefer these words by incorporating the weight of a word in the equation of PageRank as follows:

$$s(v_i) = (1-\alpha)\cdot p(v_i) + \alpha \cdot \sum_{v_j \in Adj(v_i)} \frac{w_{ji}}{\sum_{v_k \in Adj(v_j)} w_{jk}} s(v_j)$$



- Multi-word phrases are scored by using the sum of scores of individual words that comprise the phrase [Wan & Xiao(2008)].

Datasets

Dataset	#Docs	Kp	AvgKp	unigrams	bigrams	trigrams	n-grams ($n \geqslant 4$)
KDD *	834	3093	3.70	810	1770	471	42
WWW *	1350	6405	4.74	2254	3139	931	81
Nguyen	211	882	4.18	260	457	132	33

Table: A summary of our datasets

* Datasets available at `http://www.cse.unt.edu/~ccaragea/keyphrases.html`.

Baselines

- *TF-IDF.* Keyphrases are ranked based on their term-frequency-inverse document frequency score [Barker & Cornacchia(2000)].
- *ExpandRank.* A word graph was built for each paper and its local textual neighbors [Wan & Xiao(2008)].
- *TopicalPageRank (TPR).* Latent Dirichlet Allocation is used to infer the topic distribution of words and documents and keyphrases are ranked by aggregating the topic-specific scores [Liu et al.(2010)Liu, Huang, Zheng, & Sun].

Results

How does our approach compare with other existing state-of-the-art methods?

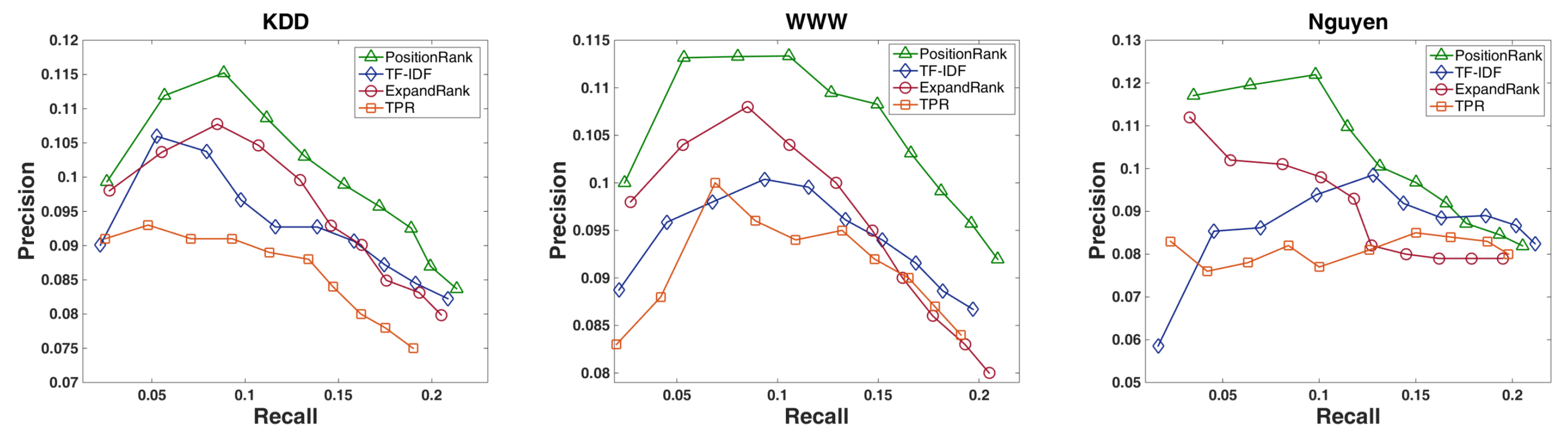


Figure: Precision-Recall curves for our proposed model and baselines on the three datasets.

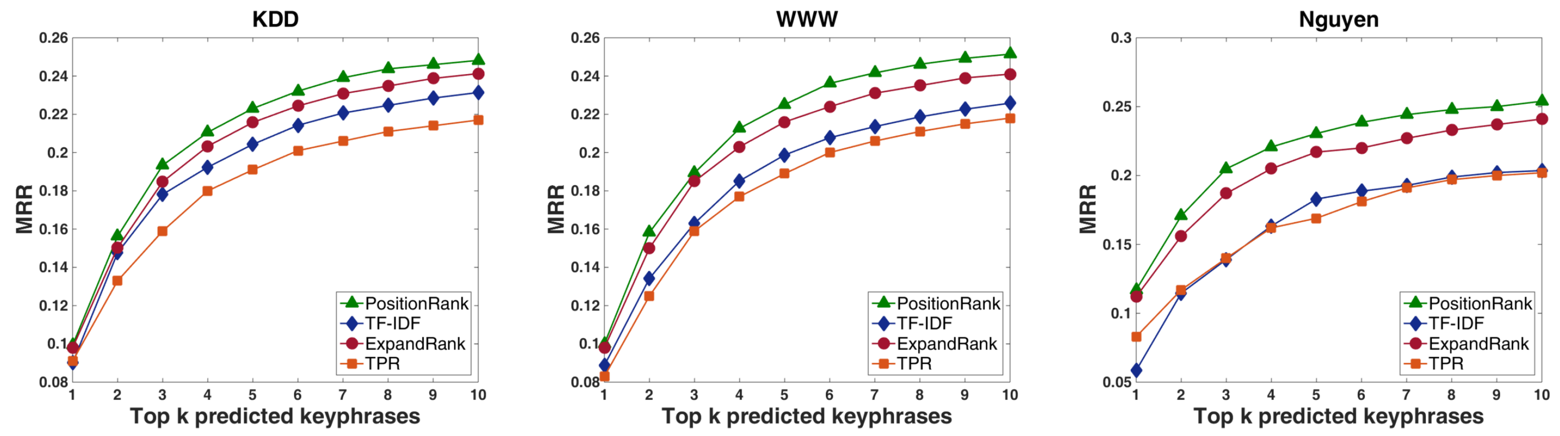


Figure: MRR curves for our proposed model and baselines on the three datasets.

Conclusion and Future Work

- Conclusions:
  - We proposed an unsupervised graph-based model which incorporates both the relative position and the frequency of a term into a biased PageRank.
  - Our experiments on three datasets show that our proposed model achieves better performance than strong baselines.
- Future directions:
  - Further evaluation of our approach on other types of documents, e.g. news articles, transcripts, etc.

References

Barker, K. & Cornacchia, N. (2000).
Using noun phrase heads to extract document keyphrases.
In Conference of the Canadian Society for Computational Studies of Intelligence, pp. 40–52.

Caragea, C., Bulgarov, F., Godea, A., & Gollapalli, S. D. (2014).
Citation-enhanced keyphrase extraction from research papers: A supervised approach.
In Proceedings of the 2014 Conferece on Empirical Methods in Natural Language Processing, pp. 1435–1446.

Gollapalli, S. D. & Caragea, C. (2014).
Extracting keyphrases from research papers using citation networks.
In Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence, pp. 1629–1635.

Haveliwala, T. H. (2002).
Topic-sensitive pagerank.
In Proceedings of the 11th international conference on World Wide Web, pp. 517–526.

Hulth, A. (2003).
Improved automatic keyword extraction given more linguistic knowledge.
In Proceedings of the 2003 conference on Empirical Methods in Natural Language Processing. pp. 216–223.

Liu, Z., Huang, W., Zheng, Y., & Sun, M. (2010).
Automatic keyphrase extraction via topic decomposition.
In Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, pp. 366–376.

Mihalcea, R. & Tarau, P. (2004).
Textrank: Bringing order into text.
In Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing.

Wan, X. & Xiao, J. (2008).
Single document keyphrase extraction using neighborhood knowledge.
In Proceedings of the Twenty-Third AAAI Conference on Artificial Intelligence, (AAAI), pp. 855–860.

Acknowledgements: This research is partially supported by a grant from the National Science Foundation to Cornelia Caragea.

WWW: https://corinaflorescu.github.io/cs/