Extracting Keyphrases from Research Papers using Citation Networks

Sujatha Das Gollapalli and Cornelia Caragea

Computer Science and Engineering, University of North Texas

Presented by: C. Lee Giles (Professor, Penn State University)

AAAI 2014

Why Keyphrase Extraction?

- Large number of scholarly documents on the Web
- The “concepts” in documents are often not provided with the documents
  - Need to be gleaned from the many details in documents.
- “Big data” times
  - Keyphrases allow for effi*cient processing of more information in less time*.

  - Keyphrases are useful in many applications such as topic tracking, information filtering and search.

Examples of Keyphrases: A snippet from the 2010 best paper award winner in the WWW conference

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. *[…]*"

- Keyphrase extraction is the task of automatically extracting descriptive phrases or concepts from a document.

Previous Approaches to Keyphrase Extraction

- Use generally only the textual content of the target document (Mihalcea and Tarau, 2004), (Liu et al., 2010).
- Wan and Xiao (2008) proposed a model that incorporates a local neighborhood of a document for extracting keyphrases.
  - Obtained improvements over models that use only textual content.
  - However, their neighborhood is limited to textually-similar documents.

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

From Data to Knowledge

A typical scientific research paper:

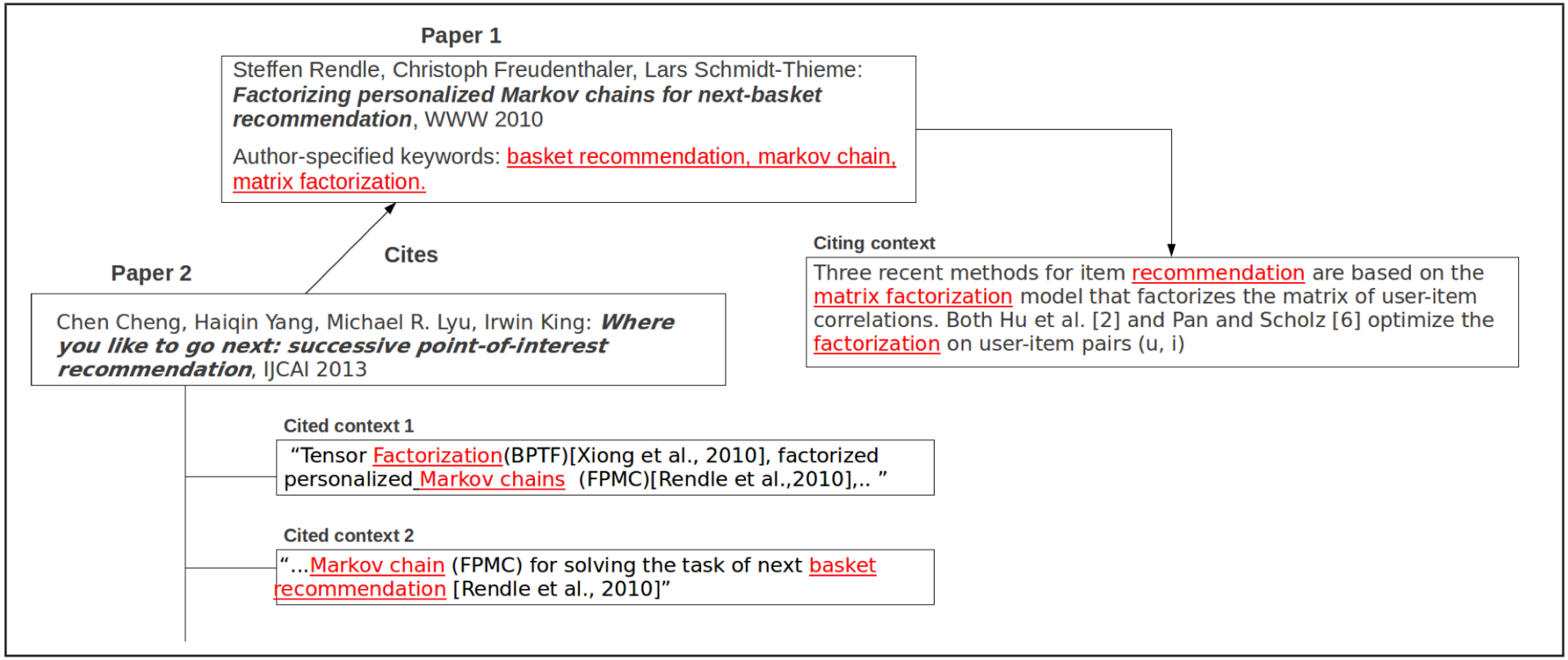
- Proposes new problems or extends the state-of-the-art for existing research problems
- Cites relevant, previously-published research papers in appropriate *contexts*.

The citations between research papers gives rise to an interlinked document network, commonly referred to as the *citation network*.

Citation Networks

- In a citation network, information flows from one paper to another via the citation relation (Shi et al, 2010)
- Citation contexts capture the influence of one paper on another as well as the flow of information
- Citation contexts or the short text segments surrounding a paper's mention serve as “micro summaries” of a cited paper!

A Small Citation Network



- *Citation contexts* are very informative!

Proposed Approach: CiteTextRank

- Citation contexts capture how one paper influences another along various aspects such as topicality, domain of study, algorithms, etc.
- How can we use these “micro summaries” in a keyphrase extraction model?

- We propose CiteTextRank: 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.

General Steps for Unsupervised Keyphrase Extraction Algorithms

1. Extract candidate words or lexical units from the textual content of the target document by applying stopword and parts-of-speech filters.
2. Score candidate words based on some criterion
   - For example, in the TFIDF scoring scheme, a candidate word score is the product of its frequency in the document and its inverse document frequency in the collection.
3. Finally, score consecutive words, phrases or *n*-grams using the sum of scores of individual words that comprise the phrase (Wan and Xiao, 2008).
4. Output the top-scoring phrases as predictions.

- CiteTextRank incorporates information from *citation contexts* while scoring candidate words in Step 2.

Graph Construction in CiteTextRank

- Let *d* be the target document and *C* be a citation network such that $d \in C$.
- Definitions:
  - A *cited context* for *d* is defined as a context in which *d is cited* by some paper d_i in the network.
  - A *citing context* for *d* is defined as a context in which *d is citing* some paper d_j in the network.
  - The content of *d* comprises its *global context*.
- Let *T* represent the types of available contexts for *d*, i.e., the *global* context of *d*, N_d^{Ctd}, the set of *cited* contexts for *d*, and N_d^{Ctg}, the set of *citing* contexts for *d*.

Graph Construction in CiteTextRank (II)

- We construct an undirected graph, $G = (V, E)$ for d as follows:
  - For each unique candidate word from all available contexts of d, add a vertex in G.
  - Add an undirected edge between two vertices v_i and v_j if the words corresponding to these vertices occur within a window of w contiguous tokens in any of the contexts.
  - The weight w_{ij} of an edge $(v_i, v_j) \in E$ is given as:

$$w_{ij} = w_{ji} = \sum_{t \in T} \sum_{c \in C_t} \lambda_t \cdot \text{cossim}(c, d) \cdot \#_c(v_i, v_j)$$

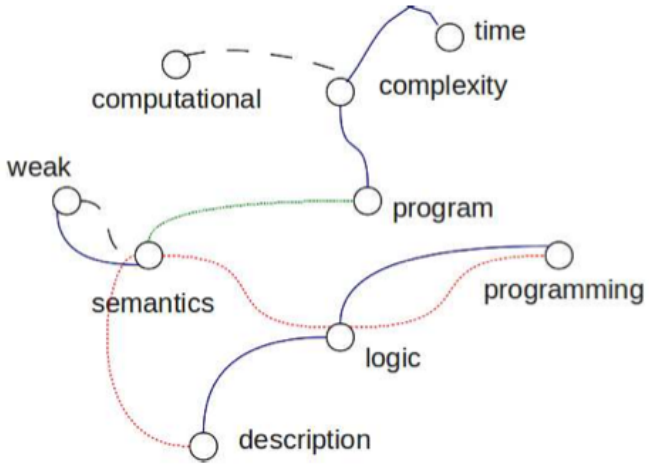
- We score vertices in G using their PageRank obtained by recursively computing:

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

(Page et al., 1999)

Parameterized Edge Weights in CiteTextRank

- Unlike simple graph edges with fixed weights, our equations correspond to *parameterized* edge weights.
- We incorporate the notion of "importance" of contexts of a certain type using the λ_t parameters.



A small word graph. Edges from different contexts are shown using different colors/line-styles.

Datasets

- We constructed three datasets of research papers and their associated citation networks using CiteSeerX. These datasets use
  1. The proceedings of the ACM Conference on Knowledge Discovery and Data Mining (KDD) and the World Wide Web Conference (WWW);
  2. The UMD dataset from Dr. Lise Getoor's research group at the University of Maryland
  3. We manually examined and annotated 100 randomly selected AAAI papers
- The author-input keyworks were used as gold-standard for evaluation.

Conference	#Titles(Org)	#Titles(CiteSeer)	#Queries	AvgKeywords	AvgCitingContexts	AvgCitedContexts
AAAI	5676	2424	93	4.15	9.77	13.95
UMD	490	439	163	3.93	20.15	34.65
WWW	2936	1350	425	4.81	15.91	17.39
KDD	1829	834	365	4.09	18.85	16.82

Table 1: Summary of datasets: #Queries represent the number of documents for which both citing, cited contexts were extracted from CiteSeerX and for which the "correct" keyphrases are available.

All datasets are available upon request.

Results

- How sensitive is CiteTextRank to its parameters?

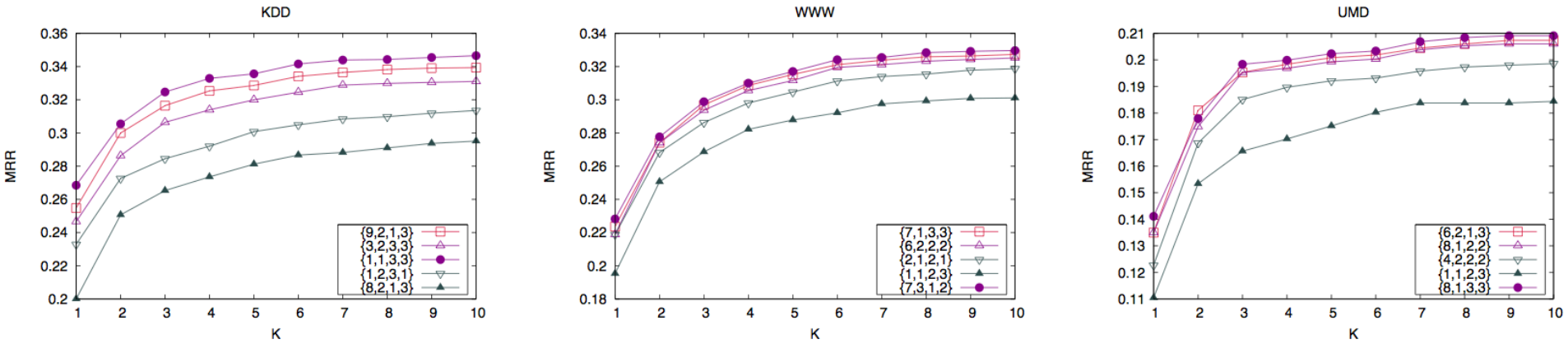


Figure: Parameter tuning for CTR. Sample configurations are shown. Setting a,b,c,d indicates window parameter is set to 'a' and the weights for content, cited and citing contexts set to 'b', 'c' and 'd', respectively.

- The varying performance of CiteTextRank with different λ_t parameters illustrates the flexibility that our model allows in treating each type of evidence differently.

Results

- How well does citation network information aid in key phrase extraction for research papers?

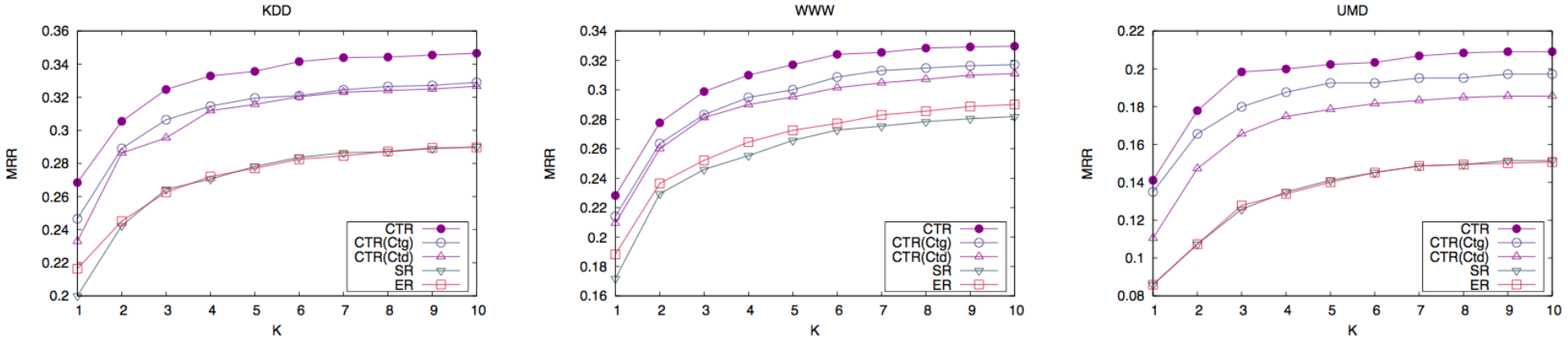


Figure: Effect of citation network information on keyphrase extraction. CTR that uses citation network neighbors is compared with ExpandRank (ER) that uses textually-similar neighbors and SingleRank (SR) that only uses the target document content.

- CiteTextRank substantially outperforms models that take into account only textually-similar documents. Cited and citing contexts contain significant hints that aid keyphrase extraction.

Results

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

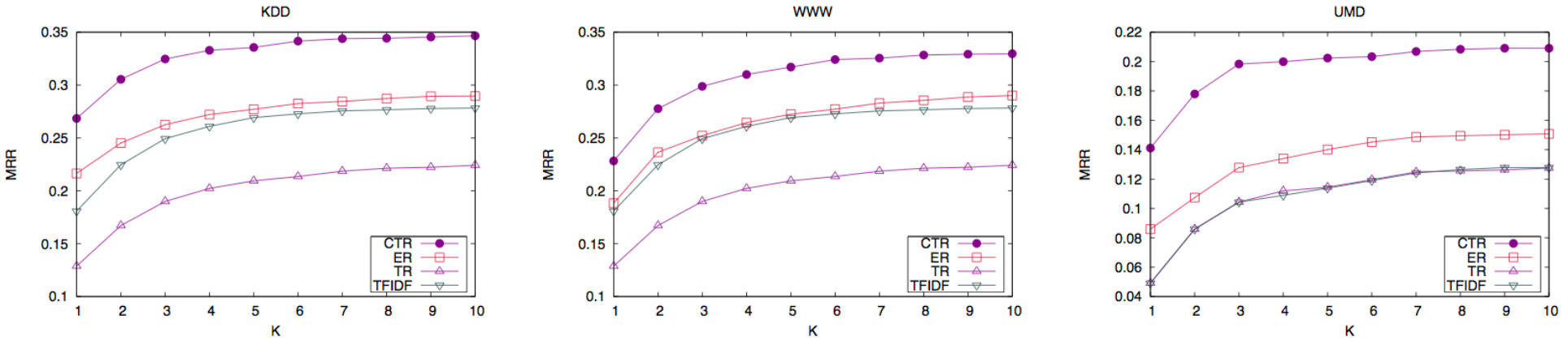


Figure: MRR curves for different keyphrase extraction methods. CiteTextRank (CTR) is compared with the baselines: TFIDF, TextRank (TR), and ExpandRank (ER).

- CiteTextRank effectively out-performs the state-of-the-art baseline models for keyphrase extraction.

Conclusions

- We proposed CiteTextRank (CTR), a flexible, unsupervised graph-based model for ranking keyphrases using multiple sources of evidence:
  - The textual content of a document and its citing and cited contexts in the interlinked document network.
- CTR gives *significant improvements* over baseline models for multiple datasets of research papers in the Computer Science domain.
- Future directions:
  - Further evaluation of CTR on other domains.
  - Extend CTR for extracting document summaries.

References

- Liu, Z., Huang, W., Zheng, Y., & Sun, M. (2010). Automatic keyphrase extraction via topic decomposition. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP ’10).
- Mihalcea, R. & Tarau, P. (2004). Textrank: Bringing order into text. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP ’04).
- Page, L., Brin, S., Motwani, R., & Winograd, T. (1999). The pagerank citation ranking: Bringing order to the web. Technical report.
- Shi, X., Leskovec, J., & McFarland, D. A. (2010). Citing for high impact. In Proceedings of the Joint Conference on Digital Libraries (JCDL ’10).
- Wan, X. & Xiao, J. (2008). Single document keyphrase extraction using neighborhood knowledge. In Proceedings of the Association for the Advancement of Artificial Intelligence (AAAI ’08).

Thank you!

Sujatha Das G.

Cornelia Caragea

C. Lee Giles

National Science Foundation
WHERE DISCOVERIES BEGIN