Learning to Extract Descriptive Keyphrases from Scholarly Big Data

Cornelia Caragea

Computer Science and Engineering
University of North Texas

June 19, 2017

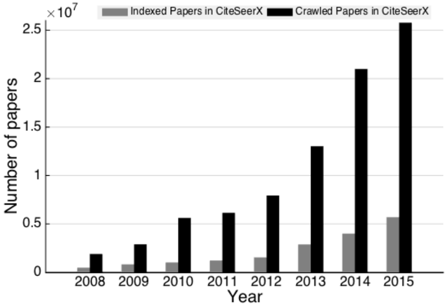
Machine Learning Group
UNT Computer Science and Engineering

Scholarly Big Data

Large number of scholarly documents on the Web

- PubMed currently has over 24 million documents
- Google Scholar is estimated to have many more million documents

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 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

- Keyphrases associated with research papers:
  - Useful in applications such as
    - topic tracking, information filtering and search, query formulation, document clustering, classification, and summarization
- However, 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 and the AAAI DL
- 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:
  - Supervised approaches [Frank et al., 1999; Turney, 2000; Hulth, 2003]
    - Formulated as binary classification, where candidate phrases are classified as either positive (i.e., keyphrases) or negative (i.e., non-keyphrases)
  - Unsupervised approaches [Mihalcea and Tarau, 2004; Wan and Xiao, 2008; Liu et al., 2010; Lahiri, Choudhury, and Caragea, 2014]
    - Formulated as a ranking problem, where candidate phrases are ranked using various measures such as tf, tf-idf, PageRank scores and other centrality measures

- 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

- 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?

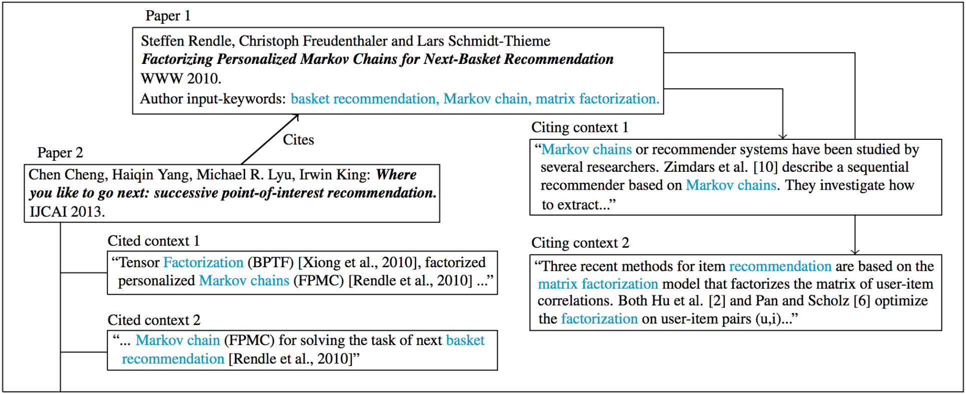
From Data to Knowledge

- 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.*
- The citations between research papers give 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, Leskovec, and McFarland, 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!

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

Citation Contexts - Previous Usage

- 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:
  - Indexing of cited papers [Ritchie, Teufel, and Robertson, 2006]
  - Author influence in document networks [Kataria, Mitra, Caragea, and Giles, 2011]
  - Scientific paper summarization [Abu-Jbara and Radev, 2011; Qazvinian, Radev, and Özgür, 2010; Qazvinian and Radev, 2008; Mei and Zhai, 2008; Lehnert et al., 1990; Nakov et al., 2004]

Citation Contexts to Keyphrase Extraction

- How can we use these contexts and how do they help in keyphrase extraction?
- We proposed:
  - **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 [Das Gollapalli and Caragea, 2014 (**AAAI**); Caragea, 2016 (**AI4DataScience**)].
  - **Citation-enhanced Keyphrase Extraction**: a supervised binary classification model built on a combination of novel features that capture information from citation contexts and existing features from previous works [Caragea et al., 2014 (**EMNLP**); Bulgarov and Caragea, 2015 (**WWW**)].

Citation Contexts to Keyphrase Extraction

- How can we use these contexts and how do they help in keyphrase extraction?
- We proposed:
  - **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 [Das Gollapalli and Caragea, 2014 (**AAAI**); Caragea, 2016 (**AI4DataScience**)].
  - **Citation-enhanced Keyphrase Extraction**: a supervised binary classification model built on a combination of novel features that capture information from citation contexts and existing features from previous works [Caragea et al., 2014 (**EMNLP**); Bulgarov and Caragea, 2015 (**WWW**)].

Unsupervised Keyphrase Extraction I

General steps for unsupervised keyphrase extraction algorithms:

1. Extract candidate words or lexical units from the content of the target document by applying stopword and parts-of-speech filters.

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**.

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.
    - MAP: 0.01; semantic: 0.3; parse: 0.05

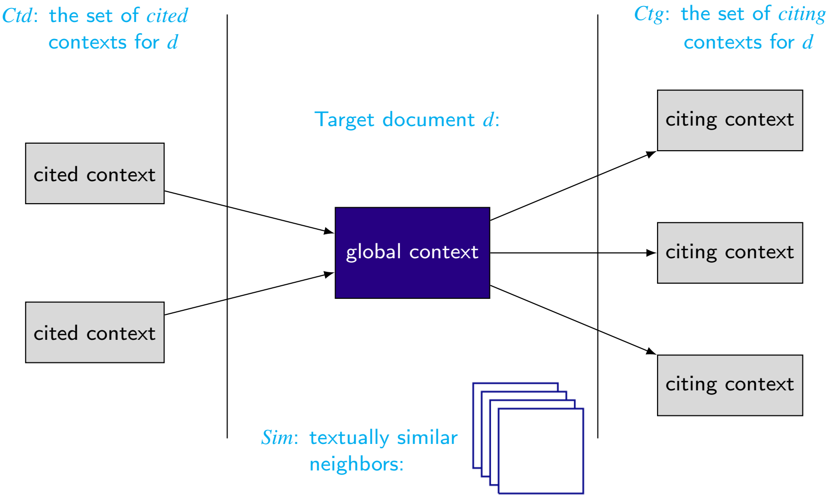
Unsupervised Keyphrase Extraction II

3. Score consecutive words, phrases or n-grams using the sum of scores of individual words that comprise the phrase [Wan and Xiao, 2008].
    - MAP semantic parse: 0.36; semantic parse: 0.35.

4. Output the top-scoring phrases as the predicted keyphrases.

CiteTextRank incorporates information from *citation contexts* while scoring candidate words in step 2, through an extension of PageRank.

CiteTextRank: Sources of Information



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

Graph Construction in CiteTextRank

We construct an undirected graph, $G = (V, E)$ for d as follows:

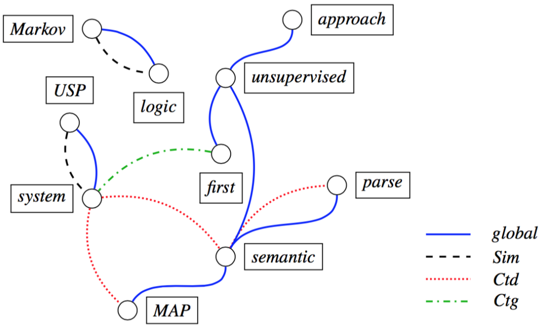
1. For each unique candidate word from all available contexts of d, add a vertex in G.
2. Add an undirected edge between two vertices v_i and v_j if the words corresponding to these vertices occur within a window of w contiguous tokens in any of the contexts.

Example Graph in CiteTextRank

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.

$w = 2$:



Parameterized Edge Weights in CiteTextRank

- The weight w_{ij} of an edge $(v_i, v_j) \in E$ is given as

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

  where λ_t is the weight for contexts of type t and C_t is the set of contexts of type $t \in T$.
- 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.

Vertex Scoring in CiteTextRank

- Initialization: $\mathbf{s} = [s(v_1), \cdots, s(v_n)] = [\frac{1}{n}, \cdots, \frac{1}{n}]$, where $n = |V|$.
- We score vertices in G using their PageRank obtained by recursively computing the equation:

$$\mathbf{s} = \alpha \cdot \widetilde{M} \cdot \mathbf{s} + (1-\alpha) \cdot \mathbf{p}, \text{where } \widetilde{m_{ij}} = \begin{cases} w_{ij} / \sum_{j=1}^{|V|} w_{ij} & \text{if } \sum_{j=1}^{|V|} w_{ij} \neq 0 \\ 0 & \text{otherwise} \end{cases}$$

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

where α is a damping factor ($\alpha = 0.85$) and $\mathbf{p} = [\frac{1}{n}, \cdots, \frac{1}{n}]$ [Page et al., 1999; Haveliwala et al., 2003]

- The PageRank score for a vertex provides a measure of its importance in the graph by taking into account global information computed recursively from the entire graph

Experiments and Results for CiteTextRank

Datasets:

- We constructed several datasets of research papers and their citation networks using CiteSeerX [Caragea et al., 2014 (**ECIR**)].
- These datasets use:
  - The proceedings of the ACM Conference on Knowledge Discovery and Data Mining (KDD) and the World Wide Web Conference (WWW);
- The author-input keyworks were used as gold-standard for evaluation.

Conference	#Docs(CiteSeerX)	#DocsUsed	AvgKeywords	AvgCtg	AvgCtd
WWW	1350	406	4.81	15.91	17.39
KDD	834	335	4.09	18.85	16.82

Table: Summary of datasets: #Queries represent the number of documents for which both citing and cited contexts were extracted from CiteSeerX and for which author-input keyphrases were available.

All datasets and code are available online.

Experimental Setting for CiteTextRank

Our experiments are organized around the following questions:

- How well does citation network information aid in keyphrase extraction for research papers?
- How does CiteTextRank perform in the absence of either citing and cited contexts?
- How does CiteTextRank compare with baseline methods?

Evaluation measures:

- Mean reciprocal rank, MRR

$$MRR = \frac{1}{|Q|} \sum_{q=1,\cdots,|Q|} \frac{1}{r_q}$$

  r_q is the rank at which the first correct prediction was found for $q \in Q$.
- Precision, Recall, F1-score.

How Well Does Citation Network Information Aid in Keyphrase Extraction for Research Papers?

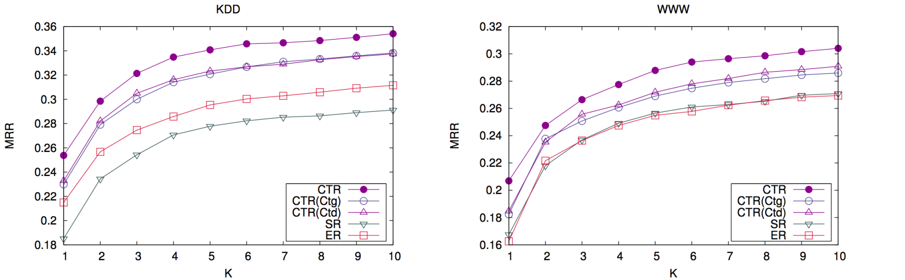
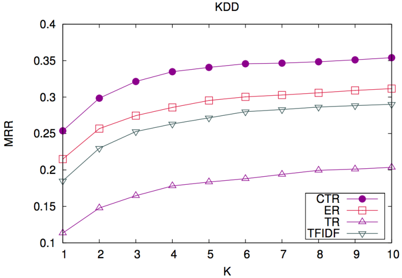


Figure: CTR that uses citation network neighbors is compared with ExpandRank (ER) that uses textually-similar neighbors and SingleRank (SR) that only uses the target document content [Wan and Xiao, 2008].

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

How Does CiteTextRank Compare with Other Existing Methods?



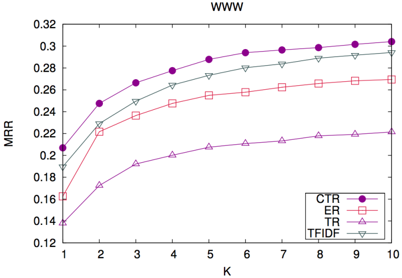


Figure: MRR curves for different keyphrase extraction methods. CTR is compared with the baselines: TFIDF, TextRank (TR) [Mihalcea and Tarau, 2004], and ExpandRank (ER) [Wan and Xiao, 2008].

CiteTextRank effectively outperforms baseline models for keyphrase extraction.

Supervised Keyphrase Extraction

- We proposed:
  - **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 [Das Gollapalli and Caragea, 2014 (**AAAI**); Caragea, 2016 (**AI4DataScience**)].
  - **Citation-enhanced Keyphrase Extraction**: a supervised binary classification model built on a combination of novel features that capture information from citation contexts and existing features from previous works [Caragea et al., 2014 (**EMNLP**); Bulgarov and Caragea, 2015 (**WWW**)].

Supervised Keyphrase Extraction - Methodology

- Generate 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.
- Represent each candidate phrase as a **vector of features.**
- Assign a positive or negative class to each phrase based on the human annotated labels.
- Use the data to train a Naïve Bayes classifier.

Features for CeKE

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

How Does CeKE Compare with Supervised 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**
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
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

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*

What Are the Most Informative Features for Keyphrase Extraction?

Rank	*Feature*	IG Score
1	*abstract tf-idf*	0.0234
2	*first position*	0.0188
3	*citation tf-idf*	0.0177
4	*relativePos*	0.0154
5	*firstPosUnder*	0.0148
6	*inCiting*	0.0129
7	*inCited*	0.0098
8	*POS*	0.0085
9	*tf-idf-Over*	0.0078

Table: Feature ranking by Information Gain on WWW.

Anecdotal Evidence

- We considered an EMNLP paper by Poon and Domingos [2009]
  - Our classifier trained on both WWW and KDD
  - We gathered from the Web 49 cited contexts and 30 citing contexts
  - The classifier was tuned to return only high-confidence keyphrases

Unsupervised Semantic Parsing$^{0.997}$

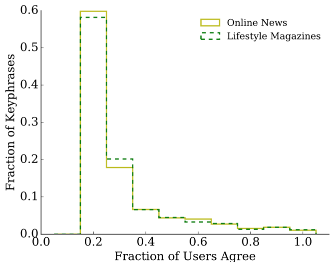
We present the first unsupervised approach to the problem of learning a **semantic parser**$^{1.000}$, using **Markov logic**$^{0.991}$. Our **USP system**$^{0.985}$ transforms dependency trees into quasi-logical forms, recursively induces lambda forms from these, and clusters them to abstract away syntactic 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.

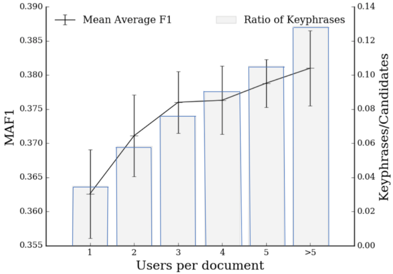
Human annotated keyphrases: *unsupervised semantic parsing, Markov logic, USP system, semantic parser*

Gray - filtered out words; *Black* - candidate phrases; **Bold cyan** - predicted keyphrases; *Numbers* - classifier's confidence

Limitations and Potential Extensions

- 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.
- Keyphrase extraction is very subjective





[Sterckx, Caragea, Demeester, Develder, 2016 (**EMNLP**)]

- Integrate terms not found in a target paper to be predicted as keyphrases (through term semantics).

Summary

- Developments in keyphrase extraction are central to ***knowledge discovery and organization*** and have a direct impact on the development of digital libraries.
- We proposed *supervised and unsupervised* models for keyphrase extraction using multiple sources of evidence
  - *Our models that integrate citation network information are state-of-the-art models to keyphrase extraction for Scholarly Data*
- We successfully extended our approaches that use citation context information to topic classification of research articles within a co-training framework [Caragea et al., 2015 (**EMNLP**)].
- We successfully leveraged knowledge from supervised and unsupervised models and designed a position-biased PageRank for KE from scholarly documents [Florescu and Caragea, 2017 (**ACL**)].

Future Directions

Extracting and Utilizing Scholarly Concept Networks

- I plan to leverage this successful work on keyphrase/concept extraction and extend it to the problem of learning semantic concept networks.
- I believe that new insights in many scientific endeavors will likely come from aggregating large amounts of digital data.
- The goal is to develop "an expert on the fly," that will continuously "read" the Scholarly Web, will discover interesting connections and hidden information between concepts, facts, or hypotheses, and will provide users with "just the right information."

An Example

- Consider the concept "teleduplication."

Definition: *teleduplication* - extracting a key's complete and precise *bitting code* at a distance via *optical decoding* and then cutting precise duplicates.
[Laxton et al., 2008]

- What should a system display for the concept "teleduplication?"

An Example

- Consider the concept "teleduplication."

Definition: *teleduplication* - extracting a key's complete and precise *bitting code* at a distance via *optical decoding* and then cutting precise duplicates.
[Laxton et al., 2008]

- What should a system display for the concept "teleduplication?"

Scholarly Knowledge Graphs

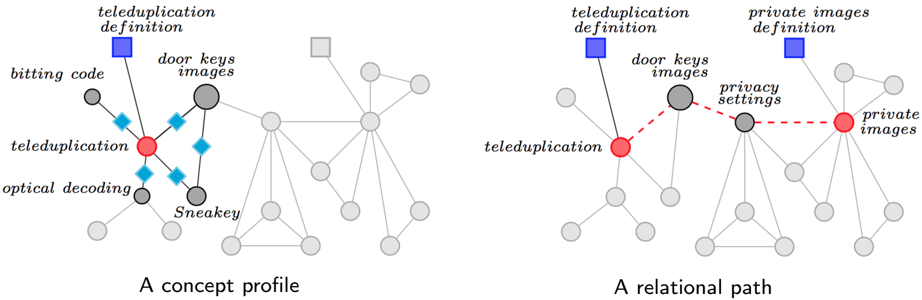


Figure: A small scholarly concept graph showing: a concept profile for "teleduplication" and one of its relational path to "private images".

- The output of this research, i.e., the concept networks, represent an initial step towards building Scholarly Knowledge Graphs with complex entities and relations.

References

- C. Florescu and C. Caragea (2017). PositionRank: An Unsupervised Approach to Keyphrase Extraction from Scholarly Documents. In: *Proceedings 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: *Proceedings of the Conference on Empirical Methods in Natural Language Processing (**EMNLP '16**).*
- C. Caragea, F. Bulgarov, R. Mihalcea (2015). Co-Training for Topic Classification of Scholarly Data. In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing (**EMNLP '15**).*
- 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**).*
- C. Caragea, F. Bulgarov, A. Godea, and S. Das Gollapalli. (2014). Citation-Enhanced Keyphrase Extraction from Research Papers: A Supervised Approach. In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing (**EMNLP '14**).*

Thank you!

- **Acknowledgements:**

National Science Foundation
WHERE DISCOVERIES BEGIN

CiteSeer$^{x}_{10M}$