Identifying Emergent Research Trends By Key Authors and Phrases

Shenhao Jiang, Animesh Prasad, Min-Yen Kan, and Kazunari Sugiyama

School of Computing, National University of Singapore

https://github.com/WING-NUS/ResearchTrends http://wing.comp.nus.edu.sg/?page_id=724

jiangshenhao@gatech.edu, {animesh, kanmy, sugiyama}@comp.nus.edu.sg

❖ Introduction

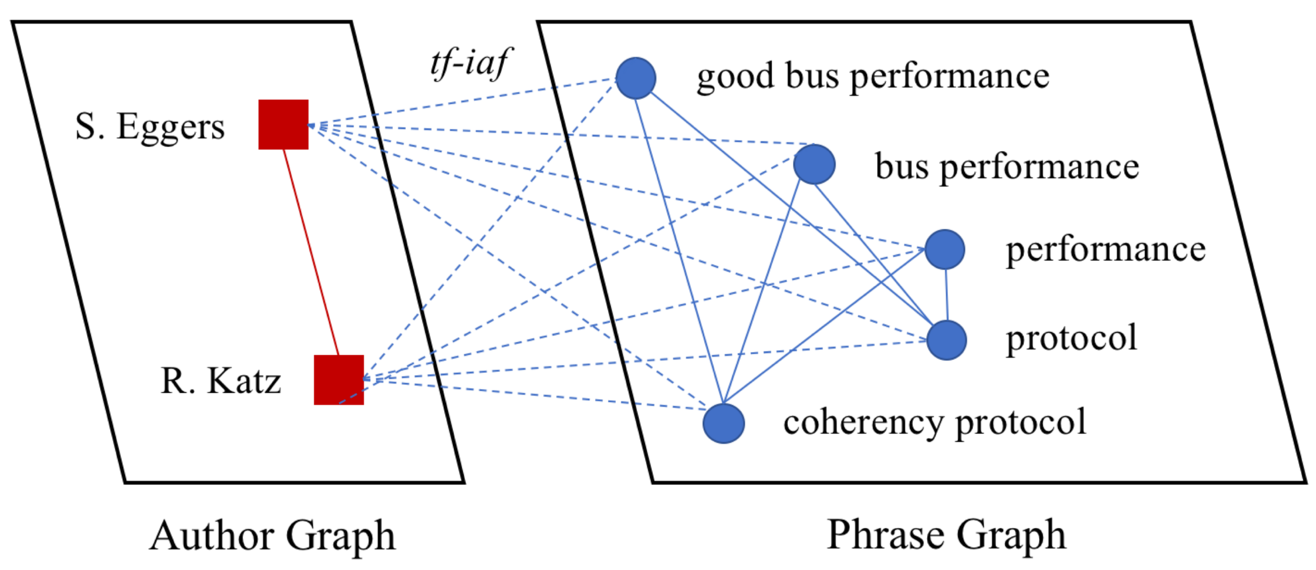
- **Motivation**:
  - Bloom of scientific publications
  - Researchers need to scan large amount of data for identification of areas with long-term impact
- **State-of-the-arts**:
  - Text Mining: LDA-type models (e.g. Dynamic Topic Models and Author Topic Model), temporal and authoring aspects of topics;
  - Citation Links: co-citation networks of papers, where tightly knit clusters represent topics, and keywords indicate trends
- **Observation:**
  - Influential authors often collaborate together
  - Important authors are more likely to write about important words which are potential trending words

❖ Proposed Techniques

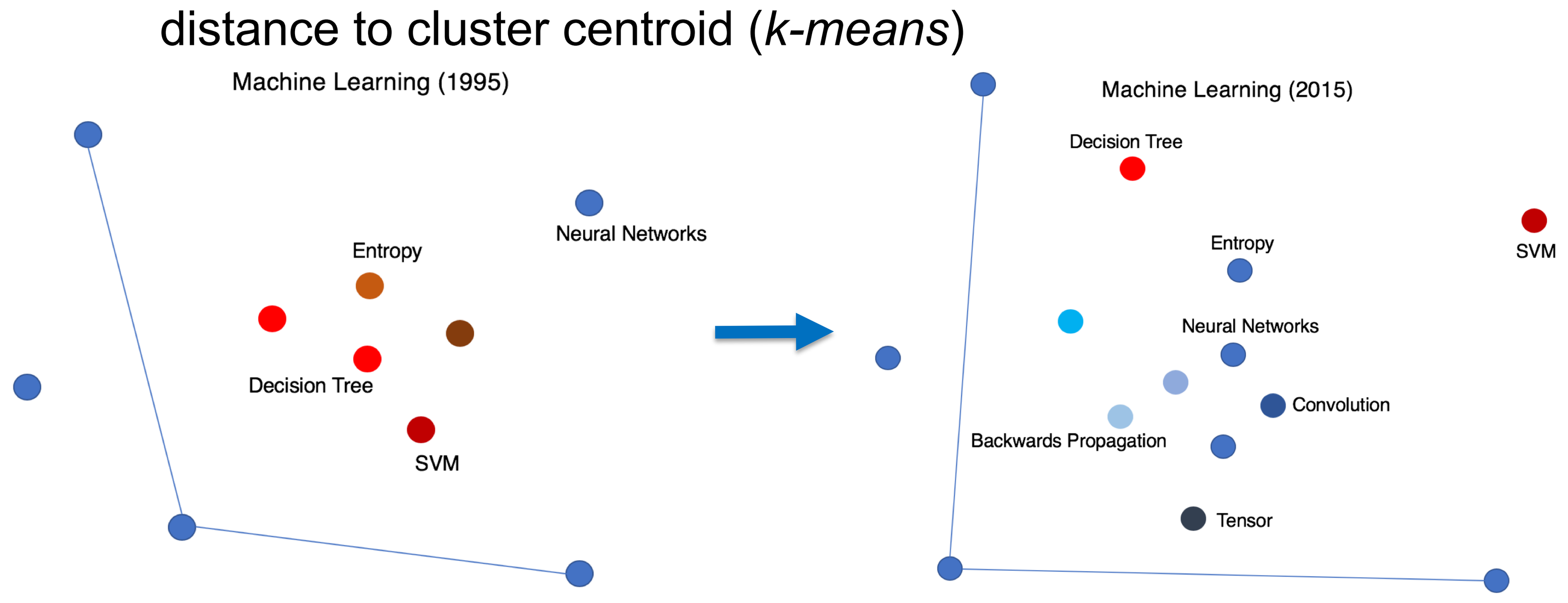
- **Step 1: MultiGraph-Ranking (MGR)**
  - Yearly grouped documents
  - Author graph and phrase graph (mutual recursion) in each year
  - Author-Author: collaboration; Phrase-Phrase: co-occurrence
  - Author-Phrase: tf-iaf

$$
\begin{aligned}
tf\text{-}iaf_{a_i,p_j} &= tf_{a_i,p_j} \times iaf_{p_j} \\
&= \frac{Occ(a_i,p_j)}{\sum_{z=1}^{n} Occ(a_i,p_z)} \times \log \frac{|A|}{|A(p_j)|},
\end{aligned}
$$

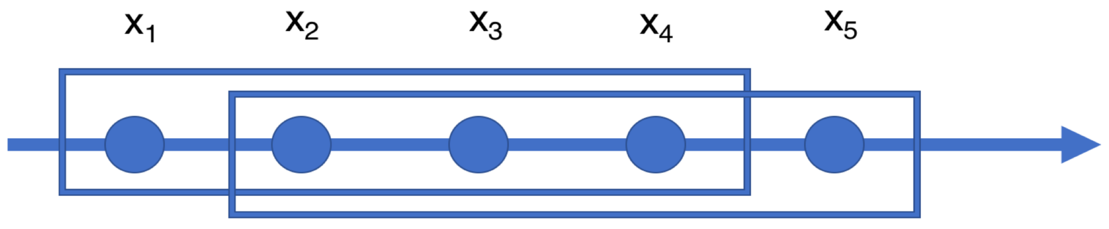
Graph Ranking with Year 1989 (Part)



- **Step 2: Word2Vec Representativeness**
  - In different timestamps, representativeness of phrases could vary drastically; therefore we enhance the score from Step 1 with the distance to cluster centroid (*k-means*)

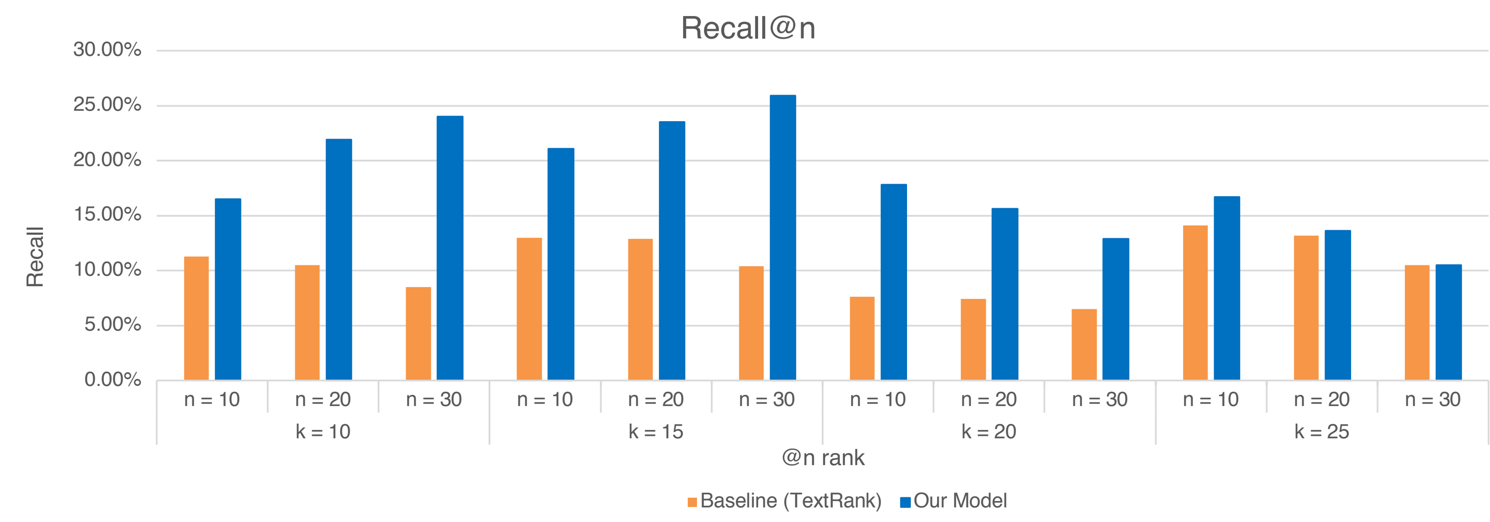


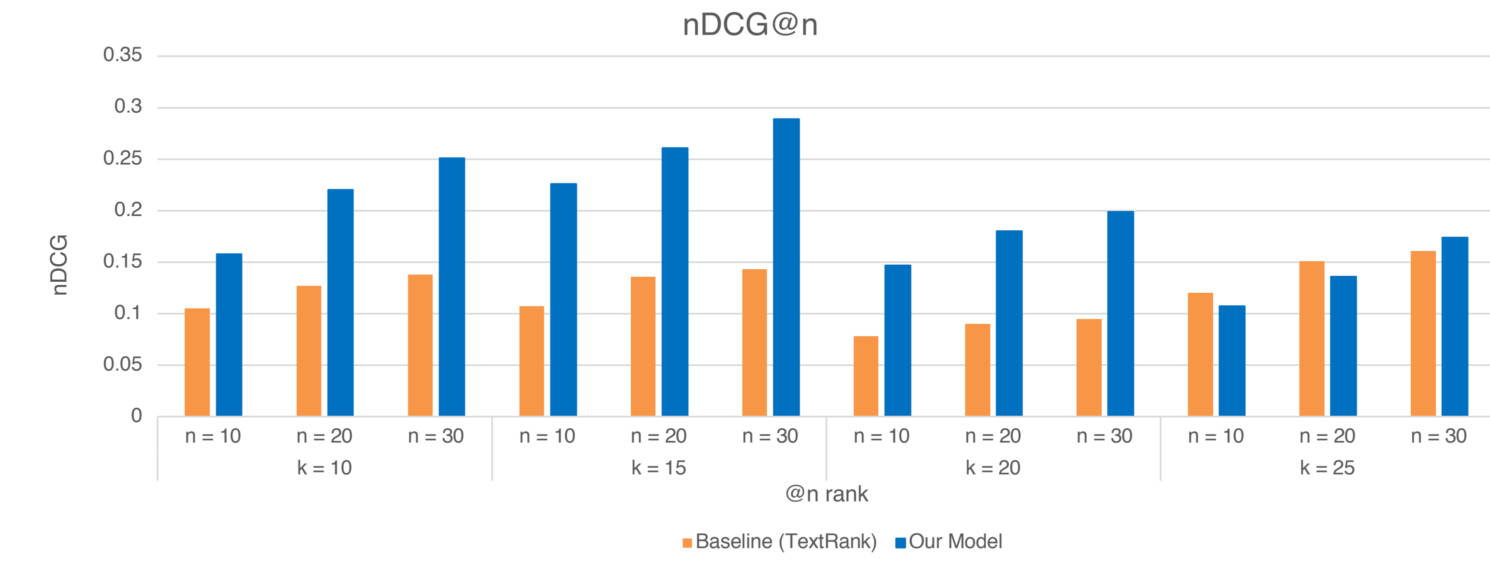
- **Step 3: RNN Predicting Scores**
  - Time series of scores: $x_1, x_2, \ldots, x_n$. We train an RNN to perform $x_{t+3} = f(x_t, x_{t+1}, x_{t+2})$ with a sliding window moving through the series.



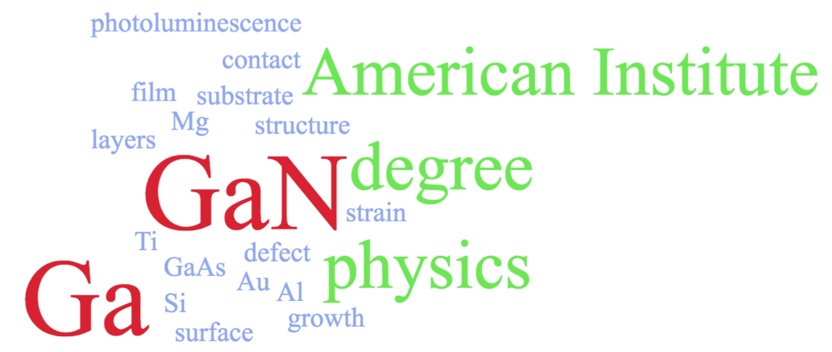
❖ Experiments & Results

- **Quantitative: ACM Periodical Dataset**
  - Abstract as doc entry
  - Field of "software engineering"
  - Baseline: replace Step 1 with standard TextRank [1]

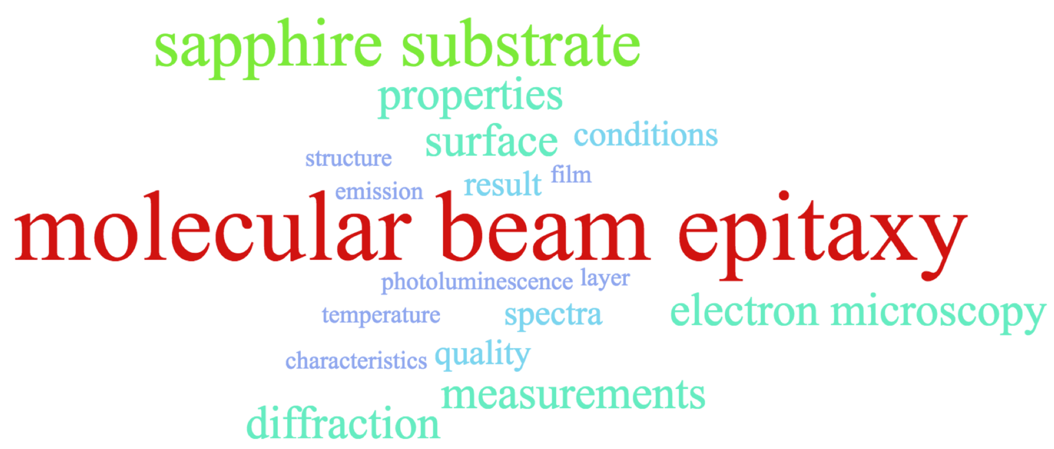




- **Qualitative: SCI & SSCI Dataset**
  - Comparative study against Shibata et al. [2]
  - Field of "Gallium Nitride (GaN)"
  - Predicting trending phrases in 2000



Baseline [2]



Our model

❖ Discussions

- Our phrase extraction model consistently outperforms the baseline TextRank, and can be taken as empirical justification for our assumption where important authors and phrases mutually influence each other
- Our extracted keyphrases work better than Shibata et al.'s work [2], and we conclude that because of the way we form phrase nodes in MGR, longer terms are compensated, and our tf-iaf concept has reduced the effects of large occurrences.
- **Future Directions**
  - Pre-train the existing Word2Vec model with our data, so there is no need to use the tf-idf average for representativeness.
  - Possible to apply to other disciplines like PubMed data, so utilize the domain experts to help us evaluate the performance.

References:

[1]: Rada Mihalcea and Paul Tarau. 2004. TextRank: Bringing Order into Texts. In Proc. of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP 2004),

[2]: Naoki Shibata, Yuya Kajikawa, Yoshiyuki Takeda, and Katsumori Matsushima. 2008. Detecting Emerging Research Fronts Based on Topological Measures in Citation Networks of Scientific Publications. Technovation