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

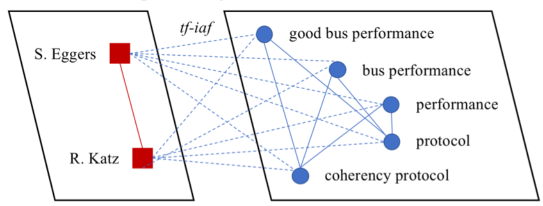
- **Motivation**:
  - Information overload in number of scientific publications
  - Users can't scan large amounts of scholarly publications to identify areas with long-term impact
- **Current State of the Art**:
  - Text Mining: adapted LDA models (e.g. Dynamic Topic Models and Author Topic Model), temporal and authoring aspects of topics;
  - Citation Links: co-citation networks of papers, where tightly knit clusters represent topics, and keywords indicate trends
- **Key Observation:**
  - Influential authors often collaborate together
  - Important authors often write words which are potentially trending

❖ Proposed Technique

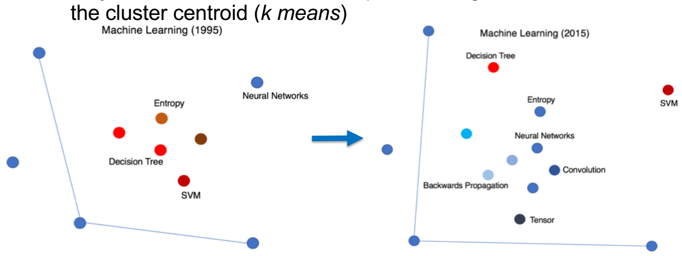
- **Step 1: Multi Graph Ranking (MGR)**
  - Group publications by year (time unit)
  - Author graph and phrase graph (mutual recursion) per year
  - Author–Author: collaboration; Phrase-Phrase: co-occurrence
  - Author–Phrase: $tf \times iaf$

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

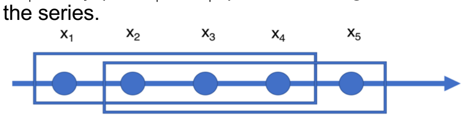
Graph Ranking with Year 1989 (Part)



- **Step 2: Word2Vec Representativeness**
  - In different timestamps, the representativeness of phrases can vary; therefore we scale the Step 1 score against the distance to the cluster centroid (*k means*)

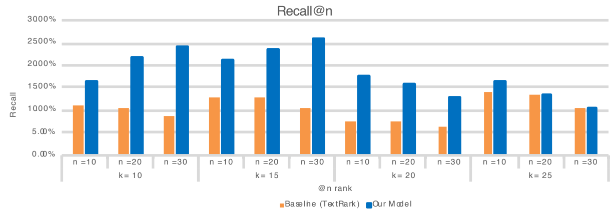


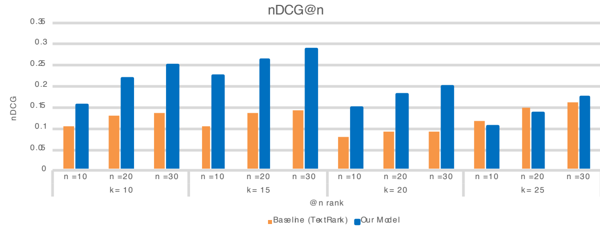
- **Step 3: RNN Predicting Scores**
  - Time series of scores: $x_1, x_2, \ldots, x_n$. We train an RNN to perform $x_{t+3} = f(x_t, x_{t+1}, x_{t+2})$ with a sliding window moving through the series.



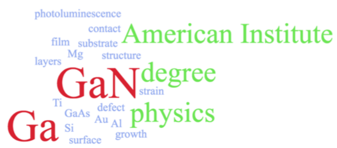
❖ Experiments & Results

- **Quantitative: ACM Periodicals**
  - Article abstract as the document unit
  - Field of "Software Engineering"
  - Baseline: replace Step 1 with standard TextRank [1]

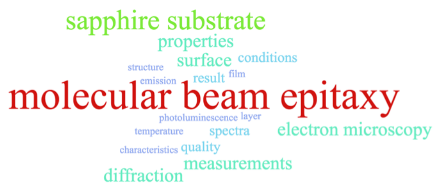




- **Qualitative: SCI & SSCI Dataset**
  - Field of Material Science on "Gallium Nitride (GaN)"
  - Predicting trending phrases in 2000



Shibata *et al.* [2]



Our model

❖ Discussion

- 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 $tf \times iaf$ 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 $tf \times idf$ average for representativeness.
  - Possible to apply to other disciplines, e.g. PubMed data, and utilize domain experts to 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