Modeling Topic-level Academic Influence in Scientific Literatures

Jiaming Shen, Zhenyu Song, Shitao Li, Zhaowei Tan, Yuning Mao, Luoyi Fu, Li Song, Xinbing Wang

Shanghai Jiao Tong University

Feb 13, 2016

Outline

1. Motivation
2. J-Index Framework
3. Reference Topic Model (RefTM)
   - Generative Model
   - Parameter Estimation
4. Experiments
   - Datasets
   - Evaluation Aspects
   - Evaluation Results
5. Conclusions & Future works

Outline

Motivation

When a beginner starts to explore a new field ...

Motivation

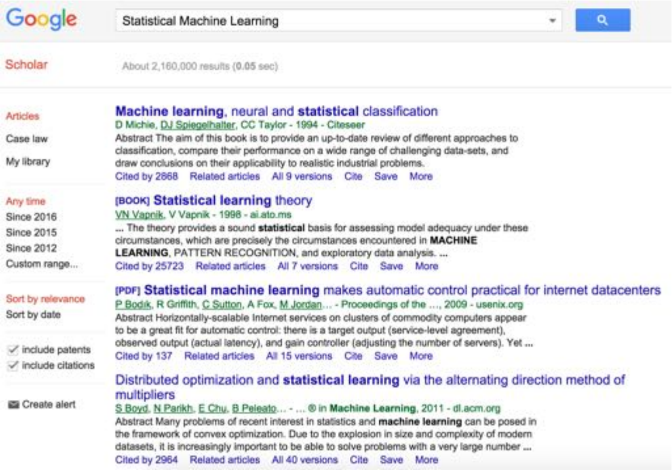

Figure 1 : Result of Google Scholar

Motivation

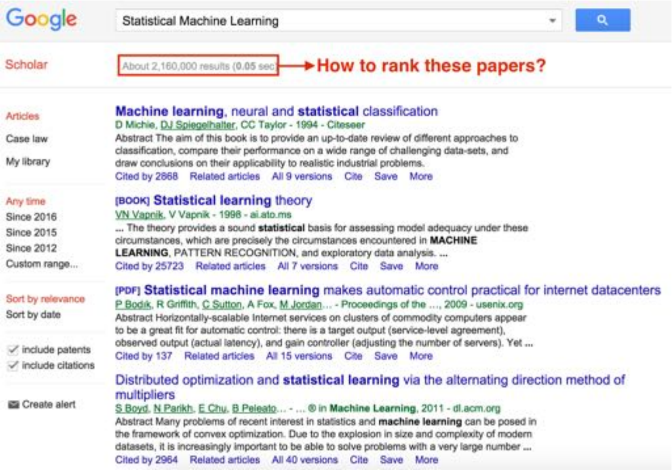

Figure 2 : Defects of Google Scholar

Motivation

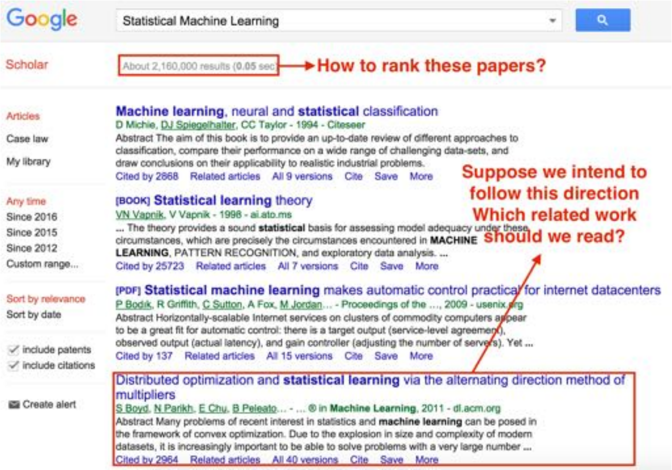

Figure 3 : Defects of Google Scholar

Motivation

Stand on the shoulders of giants

– Isaac Newton

Motivation

Find those giants !!!

Outline

Related works

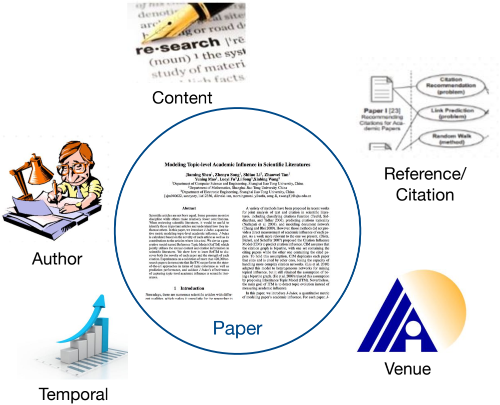

Figure 4 : Factors of one paper

Related works

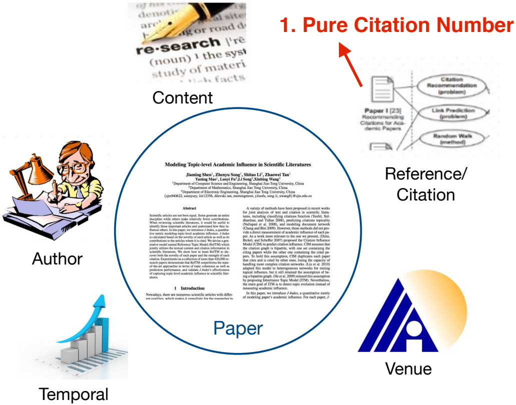

Figure 5 : Factors of one paper

Related works

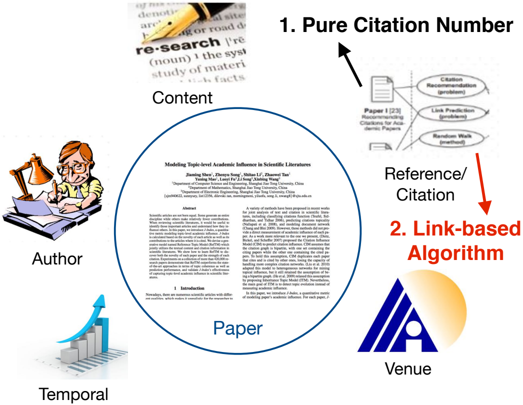

Figure 6 : Factors of one paper

Related works

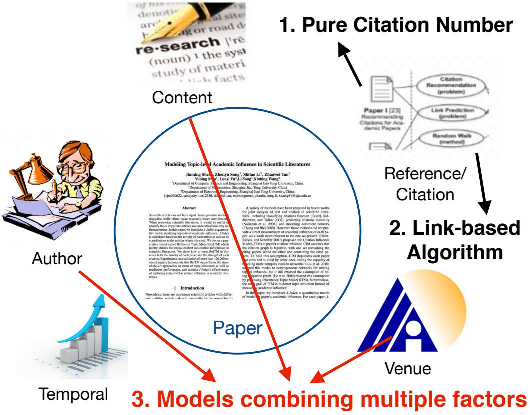

Figure 7 : Factors of one paper

Related works

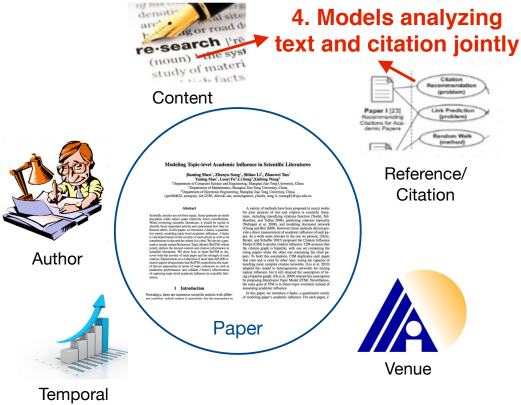

Figure 8 : Factors of one paper

Related works

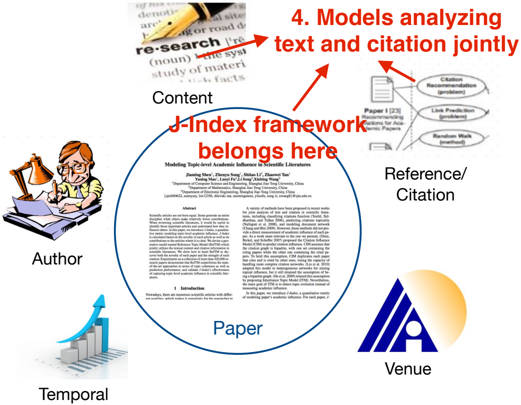

Figure 9 : Factors of one paper

J-Index Framework

- Three assumptions of J-Index:

J-Index Framework

- Three assumptions of J-Index:
  1. A paper's academic influence increases as it gains more citations.

J-Index Framework

- Three assumptions of J-Index:
  1. A paper's academic influence increases as it gains more citations.
  2. A paper with stronger citations intends to be more influential.

J-Index Framework

- Three assumptions of J-Index:
  1. A paper's academic influence increases as it gains more citations.
  2. A paper with stronger citations intends to be more influential.
  3. A paper cited by more innovative papers is more influential.

J-Index Framework

- Three assumptions of J-Index:
  1. A paper's academic influence increases as it gains more citations.
  2. A paper with stronger citations intends to be more influential.
  3. A paper cited by more innovative papers is more influential.
- We define the *J-Index* as follows:

$$\text{J-Index-Score}(u) = \sum_{c \in C(u)} \lambda(c) \times \delta(c, u)$$

- $C(u)$: the set of paper u's citations, obtained from input dataset.
- $\lambda(c)$: the innovativeness of paper c.
- $\delta(c, u)$: the citation strength between paper c and paper u.
- Both $\lambda(c)$ and $\delta(c, u)$ are obtained from subsequent model.

Outline

Generative Model

- Reference Topic Model is one way to obtain $\lambda(c)$ and $\delta(c, u)$.

Generative Model

- Reference Topic Model is one way to obtain $\lambda(c)$ and $\delta(c, u)$.
- The intuition: a researcher may write a word based on his/her own idea or "inherits" some thoughts from one of its references.

Generative Model

- Reference Topic Model is one way to obtain $\lambda(c)$ and $\delta(c, u)$.
- The intuition: a researcher may write a word based on his/her own idea or "inherits" some thoughts from one of its references.
- Topic Innovation: come from one's own idea.

Generative Model

- Reference Topic Model is one way to obtain $\lambda(c)$ and $\delta(c, u)$.
- The intuition: a researcher may write a word based on his/her own idea or "inherits" some thoughts from one of its references.
- Topic Innovation: come from one's own idea.
- Topic Inheritance: come from one of cited papers.

Generative Model

- Reference Topic Model is one way to obtain $\lambda(c)$ and $\delta(c, u)$.
- The intuition: a researcher may write a word based on his/her own idea or "inherits" some thoughts from one of its references.
- Topic Innovation: come from one's own idea.
- Topic Inheritance: come from one of cited papers.
- Citation Strength: determine which reference is selected

Generative Model

1. For each topic index $k \in \{1, \dots, K\}$
   (a) Draw a word distribution $\boldsymbol{\varphi_k} \sim \text{Dir}(\boldsymbol{\beta})$
2. For each document index $m \in \{1, \dots, M\}$
   (a) Draw a topic distribution $\boldsymbol{\theta_m} \sim \text{Dir}(\boldsymbol{\alpha})$
   (b) Draw a reference distribution $\boldsymbol{\delta_m} \sim \text{Dir}(\boldsymbol{\eta} | L_m)$
   (c) Draw an inheritance index $\lambda_m \sim \text{Beta}(\alpha_{\lambda_n}, \alpha_{\lambda_c})$
   (d) For each word $n \in \{1, \dots, N_m\}$ in document m:
      (i) Flip a coin $s_{m,n} \sim \text{Bern}(\lambda_m)$
      (ii) if $s_{m,n} = 0$:
         Draw a topic $z_{m,n} \sim \text{Multi}(\boldsymbol{\theta_m})$
         Draw a word $w_{m,n} \sim \text{Multi}(\boldsymbol{\varphi}_{z_{m,n}})$
      (iii) else ($s_{m,n} = 1$):
         Draw a reference $c_{m,n} \sim \text{Multi}(\boldsymbol{\delta_m})$
         Draw a topic $z_{m,n} \sim \text{Multi}(\boldsymbol{\theta}_{c_{m,n}})$
         Draw a word $w_{m,n} \sim \text{Multi}(\boldsymbol{\varphi}_{z_{m,n}})$

Figure 10 : Generative Model of RefTM

Generative Model

1. For each topic index $k \in \{1, \ldots, K\}$
   (a) Draw a word distribution $\boldsymbol{\varphi_k} \sim \mathrm{Dir}(\boldsymbol{\beta})$
2. For each document index $m \in \{1, \ldots, M\}$
   (a) Draw a topic distribution $\boldsymbol{\theta_m} \sim \mathrm{Dir}(\boldsymbol{\alpha})$
   (b) Draw a reference distribution $\boldsymbol{\delta_m} \sim \mathrm{Dir}(\boldsymbol{\eta}|L_m)$
   (c) Draw an inheritance index $\lambda_m \sim \mathrm{Beta}(\alpha_{\lambda_n}, \alpha_{\lambda_c})$
   (d) For each word $n \in \{1, \ldots, N_m\}$ in document m:
      (i) Flip a coin $s_{m,n} \sim \mathrm{Bern}(\boxed{\lambda_m})$ **Innovativeness**
      Topic Innovation (ii) if $s_{m,n} = 0$:
         Draw a topic $z_{m,n} \sim \mathrm{Multi}(\boldsymbol{\theta_m})$
         Draw a word $w_{m,n} \sim \mathrm{Multi}(\boldsymbol{\varphi}_{z_{m,n}})$
      Topic Inheritence (iii) else ($s_{m,n} = 1$):
         Draw a reference $c_{m,n} \sim \mathrm{Multi}(\boxed{\boldsymbol{\delta_m}})$ **Citation Strength**
         Draw a topic $z_{m,n} \sim \mathrm{Multi}(\boldsymbol{\theta}_{c_{m,n}})$
         Draw a word $w_{m,n} \sim \mathrm{Multi}(\boldsymbol{\varphi}_{z_{m,n}})$

Figure 11 : Generative Model of RefTM

Outline

1. Motivation
2. J-Index Framework
3. Reference Topic Model (RefTM)
   - Generative Model
   - Parameter Estimation
4. Experiments
   - Datasets
   - Evaluation Aspects
   - Evaluation Results
5. Conclusions & Future works

Digression: Intuition of Inference Process

- Roll a dice N times, side i shows n_i times.

Digression: Intuition of Inference Process

- Roll a dice N times, side i shows n_i times.
- What's the best estimation of each side's probability.

Digression: Intuition of Inference Process

- Roll a dice N times, side i shows n_i times.
- What's the best estimation of each side's probability.
- For each side i, probability $P_i = \frac{n_i}{N} = \frac{n_i}{\sum_{j=1}^{6} n_j}$

Digression: Intuition of Inference Process

- Roll a dice N times, side i shows n_i times.
- What's the best estimation of each side's probability.
- For each side i, probability $P_i = \frac{n_i}{N} = \frac{n_i}{\sum_{j=1}^{6} n_j}$
- Smooth Effect: Suppose we have already view each side i λ_i times.

Digression: Intuition of Inference Process

- Roll a dice N times, side i shows n_i times.
- What's the best estimation of each side's probability.
- For each side i, probability $P_i = \frac{n_i}{N} = \frac{n_i}{\sum_{j=1}^{6} n_j}$
- Smooth Effect: Suppose we have already view each side i λ_i times.
- Update: For each side i, probability $P_i = \frac{n_i + \lambda_i}{\sum_{j=1}^{6}(n_j + \lambda_j)}$

Digression: Intuition of Inference Process

- Roll a dice N times, side i shows n_i times.
- What's the best estimation of each side's probability.
- For each side i, probability $P_i = \frac{n_i}{N} = \frac{n_i}{\sum_{j=1}^{6} n_j}$
- Smooth Effect: Suppose we have already view each side i λ_i times.
- Update: For each side i, probability $P_i = \frac{n_i+\lambda_i}{\sum_{j=1}^{6}(n_j+\lambda_j)}$
- Inference: Observation $\rightarrowtail$ Parameters

Digression: Intuition of Inference Process

- Roll a dice N times, side i shows n_i times.
- What's the best estimation of each side's probability.
- For each side i, probability $P_i = \frac{n_i}{N} = \frac{n_i}{\sum_{j=1}^{6} n_j}$
- Smooth Effect: Suppose we have already view each side i λ_i times.
- Update: For each side i, probability $P_i = \frac{n_i + \lambda_i}{\sum_{j=1}^{6}(n_j + \lambda_j)}$
- Inference: Observation $\rightarrowtail$ Parameters
- In RefTM, observations: words & citations; parameters (we mainly concerned): λ and δ

RefTM Inference: Gibbs Sampling

$$p(s_i = 0 | \boldsymbol{s}_{\neg i}, \boldsymbol{w}, \boldsymbol{z}, \cdot) \propto$$
$$\frac{(n_m^{z_i(0)} - 1) + n_m^{z_i(1)} + \alpha}{n_m^{(\cdot)(0)} + n_m^{(\cdot)(1)} + K\alpha - 1} \cdot \frac{N_m^{(0)} - 1 + \alpha_{\lambda_n}}{N_m^{(1)} + (N_m^{(0)} - 1) + \alpha_{\lambda_n} + \alpha_{\lambda_c}}$$

$$p(s_i = 1 | \boldsymbol{s}_{\neg i}, \boldsymbol{w}, \boldsymbol{z}, \boldsymbol{c}_i, \cdot) \propto$$
$$\frac{n_{c_i}^{z_i(0)} + (n_{c_i}^{z_i(1)} - 1) + \alpha}{n_{c_i}^{(\cdot)(0)} + n_{c_i}^{(\cdot)(1)} + K\alpha - 1} \cdot \frac{N_m^{(1)} - 1 + \alpha_{\lambda_c}}{(N_m^{(1)} - 1) + N_m^{(0)} + \alpha_{\lambda_n} + \alpha_{\lambda_c}}$$

$$p(c_i | \boldsymbol{c}_{\neg i}, \boldsymbol{w}, \boldsymbol{z}, s_i = 1, \cdot) \propto$$
$$\frac{n_{c_i}^{z_i(0)} + (n_{c_i}^{z_i(1)} - 1) + \alpha}{n_{c_i}^{(\cdot)(0)} + n_{c_i}^{(\cdot)(1)} + K\alpha - 1} \cdot \frac{R_m^{c_i} - 1 + \eta}{R_m^{(\cdot)} + L_m \eta - 1}$$

$$p(z_i | \boldsymbol{z}_{\neg i}, \boldsymbol{w}, s_i = 0, \cdot) \propto$$
$$\frac{n_{z_i}^{w_i} + \beta - 1}{n_{z_i}^{(\cdot)} + V\beta - 1} \cdot \frac{(n_m^{z_i(0)} - 1) + n_m^{z_i(1)} + \alpha}{n_m^{(\cdot)(0)} + n_m^{(\cdot)(1)} + K\alpha - 1}$$

$$p(z_i | \boldsymbol{z}_{\neg i}, \boldsymbol{w}, s_i = 1, c_i, \cdot) \propto$$
$$\frac{n_{z_i}^{w_i} + \beta - 1}{n_{z_i}^{(\cdot)} + V\beta - 1} \cdot \frac{n_{c_i}^{z_i(0)} + (n_{c_i}^{z_i(1)} - 1) + \alpha}{n_{c_i}^{(\cdot)(0)} + n_{c_i}^{(\cdot)(1)} + K\alpha - 1}$$

Algorithm 1 Gibbs Sampling Algorithm for RefTM

Input: $K, \boldsymbol{w}, \alpha, \beta, \eta, \lambda_c, \lambda_n$
Output: Parameter sets $\{\boldsymbol{\theta}, \boldsymbol{\varphi}, \boldsymbol{\delta}, \boldsymbol{\lambda}\}$
Read in data and zero out all count caches
Randomly initialize $\boldsymbol{z}_i, \boldsymbol{c}_i, \boldsymbol{s}_i$
for $iter = 1$ to N_{iter} **do**
  for all documents $m \in [1, M]$ **do**
    for all words $n \in [1, N_m]$ in document m **do**
      if $s_{m,n} = 0$ **then**
        Update the counts $n_m^{(k)(0)}, n_m^{(0)}$
      else
        Update the counts $n_c^{(k)(1)}, n_c^{(1)}, R_m^c, R_m$
      Draw a new $\tilde{s}$ from Eqs.(2-3)
      if $\tilde{s} = 0$ **then**
        Update the counts $n_k^{w_{m,n}}, n_k$
        Draw a new topic $\tilde{k}$ from Eq.(5)
        Update the counts $n_m^{(\tilde{k})(0)}, n_m^{(0)}, n_{\tilde{k}}^{w_{m,n}}, n_{\tilde{k}}$
      else
        Draw a new reference $\tilde{c}$ from Eq.(4)
        Update the counts $R_m^{\tilde{c}}, R_m, n_k^{w_{m,n}}, n_k$
        Draw a new topic $\tilde{k}$ from Eq.(6)
        Update the counts $n_{\tilde{c}}^{(\tilde{k})(1)}, n_{\tilde{c}}, n_{\tilde{k}}^{w_{m,n}}, n_{\tilde{k}}$
Read out parameters set $\boldsymbol{\theta}, \boldsymbol{\varphi}, \boldsymbol{\lambda}, \boldsymbol{\delta}$ by Eqs.(7-10)

Figure 12 : Gibbs sampling equations & Algorithm for RefTM

Visualization of RefTM's output

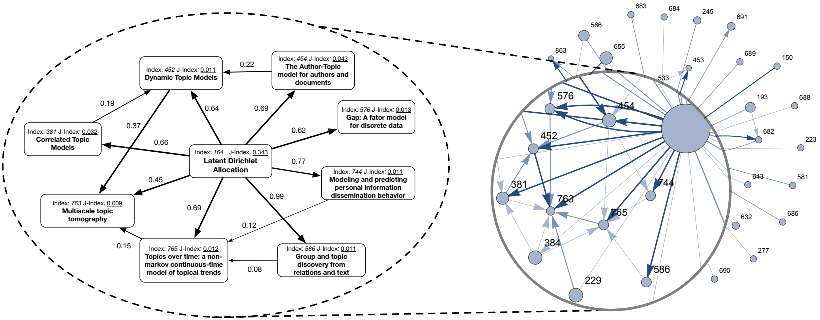

Figure 13 : Right hand side is an illustrative citation graph in which the thickness of edge represents the citation strength and the vertex size indicates one papers academic influence. Left hand side presents each paper's *J-Index*.

Outline

Datasets

- Dataset 1: a large unsupervised collection of 426728 articles with over 209 million citations.
- Dataset 2: a small supervised collection of 799 papers obtained from (Liu et al. 2010).
- The average paper length of two corpora are 83 and 98 words.

Evaluation Aspects

- Topic Coherence

Evaluation Aspects

- Topic Coherence
  1. Metrics: PMI-Score (Newman et al. 2010) and *topic coherence-Score* (Mimno et al. 2011).

Evaluation Aspects

- Topic Coherence
  1. Metrics: PMI-Score (Newman et al. 2010) and *topic coherence-Score* (Mimno et al. 2011).
  2. Using dataset 1 and an external dataset of 3.34 million papers when calculating PMI-Score.

Evaluation Aspects

- Topic Coherence
  1. Metrics: PMI-Score (Newman et al. 2010) and *topic coherence-Score* (Mimno et al. 2011).
  2. Using dataset 1 and an external dataset of 3.34 million papers when calculating PMI-Score.
- Citation Strength Prediction

Evaluation Aspects

- Topic Coherence
  1. Metrics: PMI-Score (Newman et al. 2010) and *topic coherence-Score* (Mimno et al. 2011).
  2. Using dataset 1 and an external dataset of 3.34 million papers when calculating PMI-Score.
- Citation Strength Prediction
  1. Using dataset 2, in which the strength of each citation is classified into three levels – "1, 2, 3".

Evaluation Aspects

- Topic Coherence
  1. Metrics: PMI-Score (Newman et al. 2010) and *topic coherence-Score* (Mimno et al. 2011).
  2. Using dataset 1 and an external dataset of 3.34 million papers when calculating PMI-Score.
- Citation Strength Prediction
  1. Using dataset 2, in which the strength of each citation is classified into three levels – "1, 2, 3".
  2. Metrics: averaged AUC value for decision boundaries "1 vs. 2, 3" and "1, 2 vs. 3".

Evaluation Aspects

- Topic Coherence
  1. Metrics: PMI-Score (Newman et al. 2010) and *topic coherence-Score* (Mimno et al. 2011).
  2. Using dataset 1 and an external dataset of 3.34 million papers when calculating PMI-Score.
- Citation Strength Prediction
  1. Using dataset 2, in which the strength of each citation is classified into three levels – "1, 2, 3".
  2. Metrics: averaged AUC value for decision boundaries "1 vs. 2, 3" and "1, 2 vs. 3".
- Case Study: Rank INFOCOM

Topic Coherence

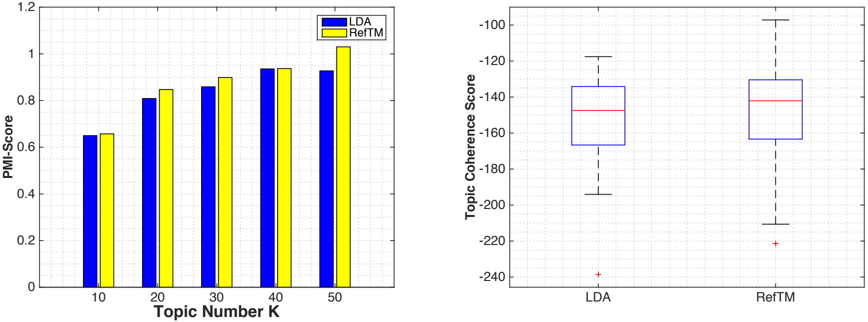

Figure 14 : Topic Coherence Evaluation

- PMI-Score: RefTM outperforms LDA by 12% when $K = 50$.
- *Topic Coherence-Score*: RefTM outperforms LDA slightly.

Citation Strength Prediction

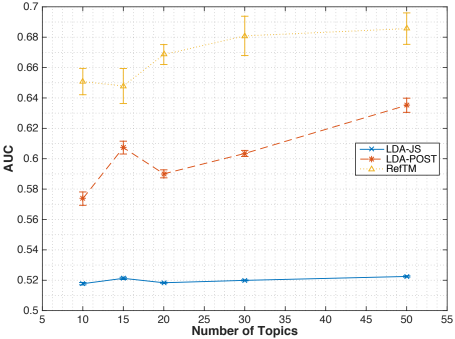

Figure 15 : Citation Strength Prediction measured by averaged AUC

- Reduce the normalization constraint of δ in RefTM.
- RefTM clearly outperforms two baseline methods.

Case Study: Rank INFOCOM

Table 2: Top 5 Articles in INFOCOM 2003 randked by *J-Index* & citaions

Title	*J-Index*	citation counts
Top 5 Articles in INFOCOM 2003 ranked by *J-Index*		
Ad hoc positioning system (APS) using AOA	6.75	115
Performance anomaly of 802.11b	5.17	127
Packet leashes: a defense against wormhole attacks in wireless networks	4.13	74
Unreliable sensor grids: coverage, connectivity and diameter	4.00	82
Sensor deployment and target localization based on virtual forces	3.61	60
Top 5 Articles in INFOCOM 2003 ranked by citation number		
Performance anomaly of 802.11b	5.17	127
Ad hoc positioning system (APS) using AOA	6.75	115
Optimal routing, link scheduling and power control in multihop wireless networks	2.26	109
Sprite: a simple, cheat-proof, credit-based system for mobile ad-hoc networks	2.43	88
Unreliable sensor grids: coverage, connectivity and diameter	4.00	82

Figure 16 : Citation Strength Prediction measured by averaged AUC

- Rankings by *J-Index* and citations number are correlated.
- *J-Index* favors those paper that propose novel "ideas".

Outline

Conclusions & Future works

- Conclusions:

Conclusions & Future works

- Conclusions:
  1. Model academic influence – facilitate ranking and recommendation.

Conclusions & Future works

- Conclusions:
  1. Model academic influence – facilitate ranking and recommendation.
  2. J-Index framework – consider citation strength and paper's novelty.

Conclusions & Future works

- Conclusions:
  1. Model academic influence – facilitate ranking and recommendation.
  2. J-Index framework – consider citation strength and paper's novelty.
  3. Reference Topic Model – combine citation network into topic model.

Conclusions & Future works

- Conclusions:
  1. Model academic influence – facilitate ranking and recommendation.
  2. J-Index framework – consider citation strength and paper's novelty.
  3. Reference Topic Model – combine citation network into topic model.
- Future works:

Conclusions & Future works

- Conclusions:
  1. Model academic influence – facilitate ranking and recommendation.
  2. J-Index framework – consider citation strength and paper's novelty.
  3. Reference Topic Model – combine citation network into topic model.
- Future works:
  1. RefTM in the incremental citation network.

Conclusions & Future works

- Conclusions:
  1. Model academic influence – facilitate ranking and recommendation.
  2. J-Index framework – consider citation strength and paper's novelty.
  3. Reference Topic Model – combine citation network into topic model.
- Future works:
  1. RefTM in the incremental citation network.
  2. Consider multiple factors, especially the temporal information.

References

- Dietz, L.; Bickel, S.; and Scheffer, T. 2007. Unsupervised prediction of citation influences. In *Proceedings of the 24th international conference on Machine learning*, 233 – 240. ACM.
- Liu, L.; Tang, J.; Han, J.; Jiang, M.; and Yang, S. 2010. Mining topic-level influence in heterogeneous networks. In *Proceedings of the 19th ACM international conference on Information and knowledge management*, 199 – 208. ACM.
- Mimno, D.; Wallach, H. M.; Talley, E.; Leenders, M.; and McCallum, A. 2011. Optimizing semantic coherence in topic models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 262 – 272. ACM.
- Newman, D.; Lau, J. H.; Grieser, K.; and Baldwin, T. 2010. Automatic evaluation of topic coherence. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, 100 – 108. ACM.

Thank you!

Q & A