A Neural Probabilistic Model for Context Based Citation Recommendation

Wenyi Huang†, Zhaohui Wu‡, Chen Liang†, Prasenjit Mitra†‡, C. Lee Giles†‡

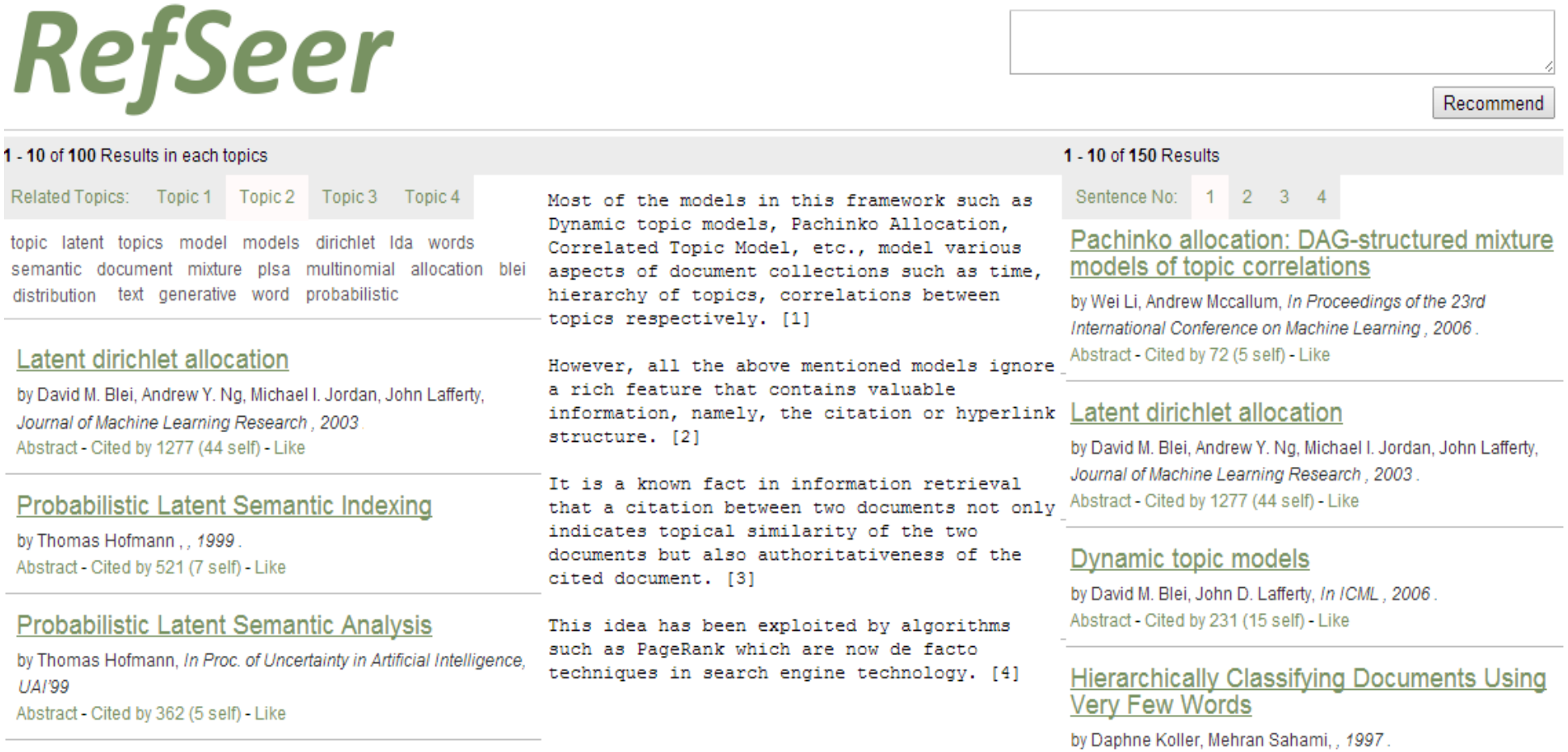
†Information Sciences and Technology, ‡Computer Sciences and Engineering
The Pennsylvania State University
University Park, PA 16802
{harrywy,laowuz}@gmail.com
{cul226,pmitra,giles}@ist.psu.edu

RefSeer

CiteSeerx β

Problem Definition

- What is citation recommendation?
- Two Types of Recommendation: 1) global 2)local



Our Contribution

- We propose a neural probabilistic model that learns the probability of citing a paper given a citation context based on distributed representations of words and documents.

- The model is implemented in RefSeer, ,a citation recommendation engine, for public uses.

- The model was trained and evaluated on the entire CiteSeer dataset which consists of 10,760,318 pairs of citation contexts and cited documents from 1,017,457 papers.

- Compared to other state-of-the-art context-based methods, our model shows significant improvement on various performance metrics, with a 5% gain in recall@10, a 2% gain in MRR and MAP, and a 3% gain in nDCG.

Motivation of this Work

- Authors use different words and phrases when citing the same article.
- The semantics of the cited documents should be close to the citation contexts.

Citation Context Examples	Cited Paper
1. ...For example, PageRank [*] can be applied to the hyperlink structure on domains to obtain domain rank scores... 2. ..., and the unique solution vector r(i) can be expressed as the eigenvector of a matrix [*] or as the stationary probability of a random walk... 3. ... Rank sinks [*] are defined to be a set of nodes which have links between themselves but no links to the other nodes. 4. There is a lot of research work on static information network analysis, including ... , and node ranking [*, *], ...	L. Page, S. Brin, R. Motwani, and T. Winograd. 1998. *The Pagerank Citation Ranking: Bringing Order to the Web.* Technical report, Stanford Digital Library Technologies Project, 1998.

Distributed Representation

- One hit representation:

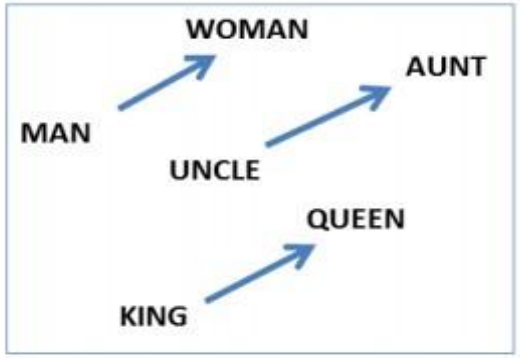
  movie [0 0 0 0 0 0 0 0 1 0 0 0 0] && film [0 0 0 0 0 1 0 0 0 0 0 0 0] = 0

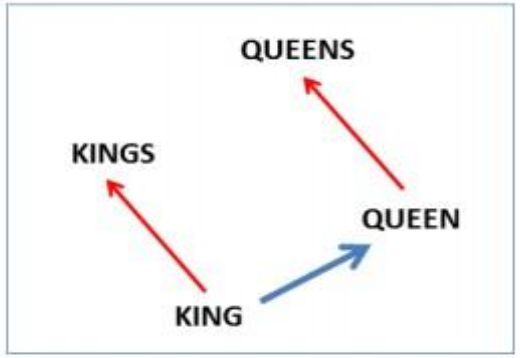
- Distributional similarity based representations
  - Soft clustering word representations
    - LSA/LDA
  - Neural network based distributed representations:
    - Words are projected into n-dimensional continues-valued vectors (n= 50, 100,…)

```
news 0.0937356086753 0.747945120724 -0.0750619884446 -0.226063114311 -1.00569231962 0.516947986429
price 0.287862881268 0.625123596091 0.382014892929 -0.0966298530516 -1.19724365222 0.0196462314248
very 0.323354305535 0.321417782731 0.483145963078 -0.157390738802 -0.962555596754 0.00207746146949
industry 0.0291885988722 0.504720529139 -0.122992000712 0.391162080274 -1.20667627561 0.05815952581
sales 0.141502071535 0.627457134652 0.144504638422 0.00177222515462 -0.971236693536 0.0857220496627
second 0.251871859833 0.65748186686 0.152971131984 -0.350360474704 -0.935662471132 0.167607205076 0
so 0.437489708071 0.637964913163 0.366912669659 -0.049045606276 -0.466581719757 0.373365386391 0.40
```

Distributed Representation

- These representations are good at capturing syntactic and semantic regularities in language, and that each relationship is characterized by a relation-specific vector offset.





$$Vector(Man) - Vector(Woman) \approx Vector(King) - Vector(Queen)$$

$$Vector(Kings) - Vector(King) \approx Vector(Queens) - Vector(Queen)$$

Turian et al. *Word representations: A simple and general method for semi-supervised learning*. ACL, 2010.

Tomas Mikolov, et al. *Linguistic Regularities in Continuous Space Word Representations*. NAACL 2013.

Problem Formulation

- Model the citation context given the cited paper $p(c|d)$. Assume that we have $|C|$ pairs of citation context c_t and cited document d_t. The objective is to maximize the log-likelihood:

$$\text{Maximize} \quad \sum_{t=1}^{|C|} \log p(c_t|d_t)$$

where $p(c_t|d_t) = p\left(w_{t_1}, \cdots, w_{t_{|c_t|}} \middle| d_t\right) = \prod_{i=1}^{|c_t|} p(w_i|d_t)$

- The objective function can be written as:

$$\text{Maximize} \quad \sum_{t=1}^{|C|} \sum_{i=1}^{|c_t|} \log p\left(w_{t_i} \middle| d_t\right)$$

Neural Probabilistic Model

- $p(w|d)$ can be defined using a softmax function:

$$p(w|d) = \frac{\exp(s_\theta(w,d))}{\sum_{i=1}^{|V|} \exp(s_\theta(w_i,d))}$$

where $|V|$ is the size of the word vocabulary.

- $s_\theta(\cdot)$ is the neural network scoring(activate) function. w and d are projected into n-dimensional continues-valued vectors v_w and v_d:

$$s_\theta(w,d) = f(v_w, v_d)$$

where $f(v_w, v_d) = \frac{1}{1+\exp(-v_w \cdot v_d)}$

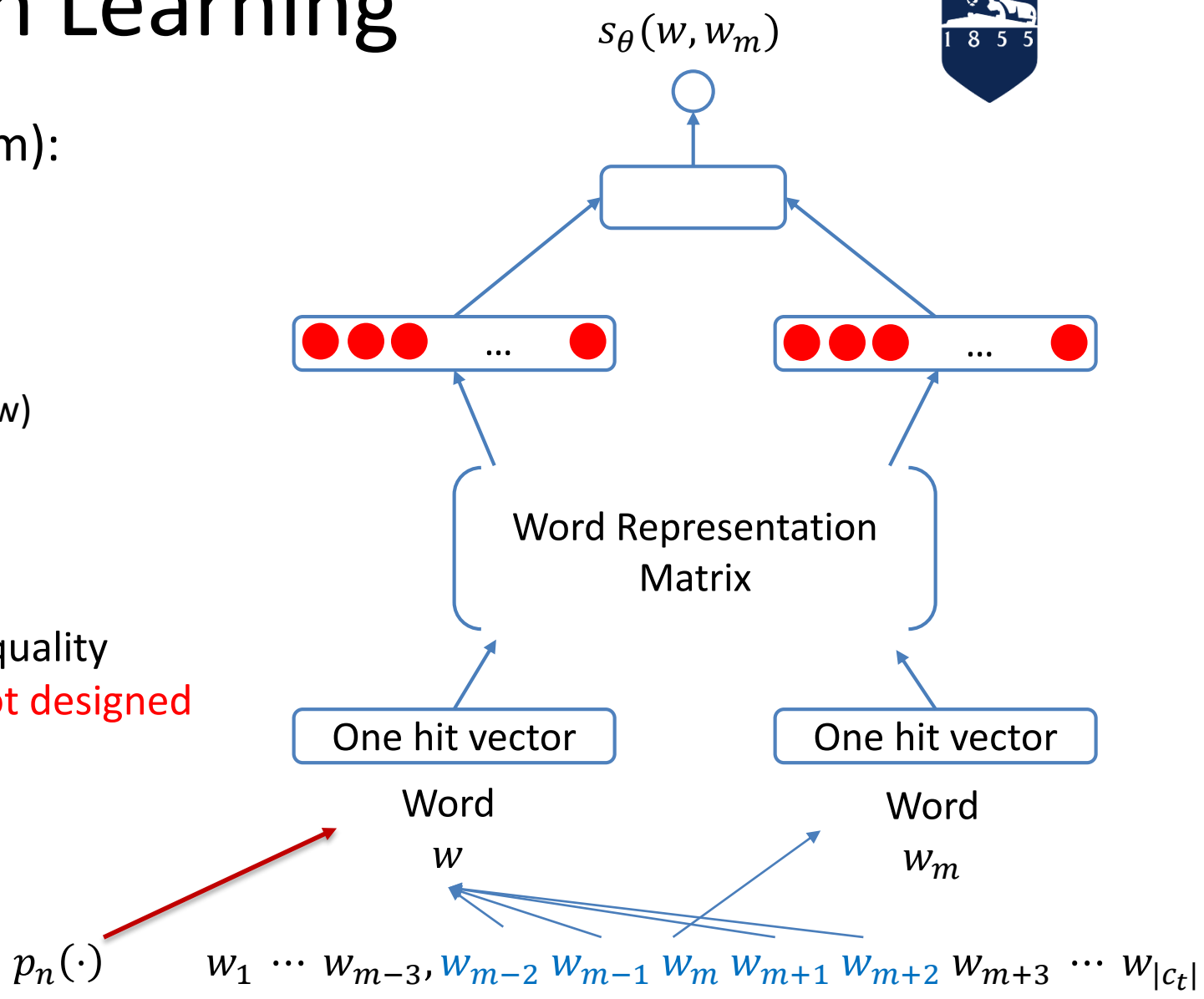
Word Representation Learning

- Google's word2vec model (skip-gram):

$$\ell_m(\theta) = \log s_\theta(w_p, w_m) + \sum_i^k \left[\log\left(1 - s_\theta(w_n, w_m)\right)\right]$$

  - w_p: positive examples (words in the window)
  - w_n: negative example generated by $p_n(\cdot)$
  - k: number of negative examples

- The skip-gram model is able to learn high quality representations of words. However, it is not designed to estimated the conditional probability.

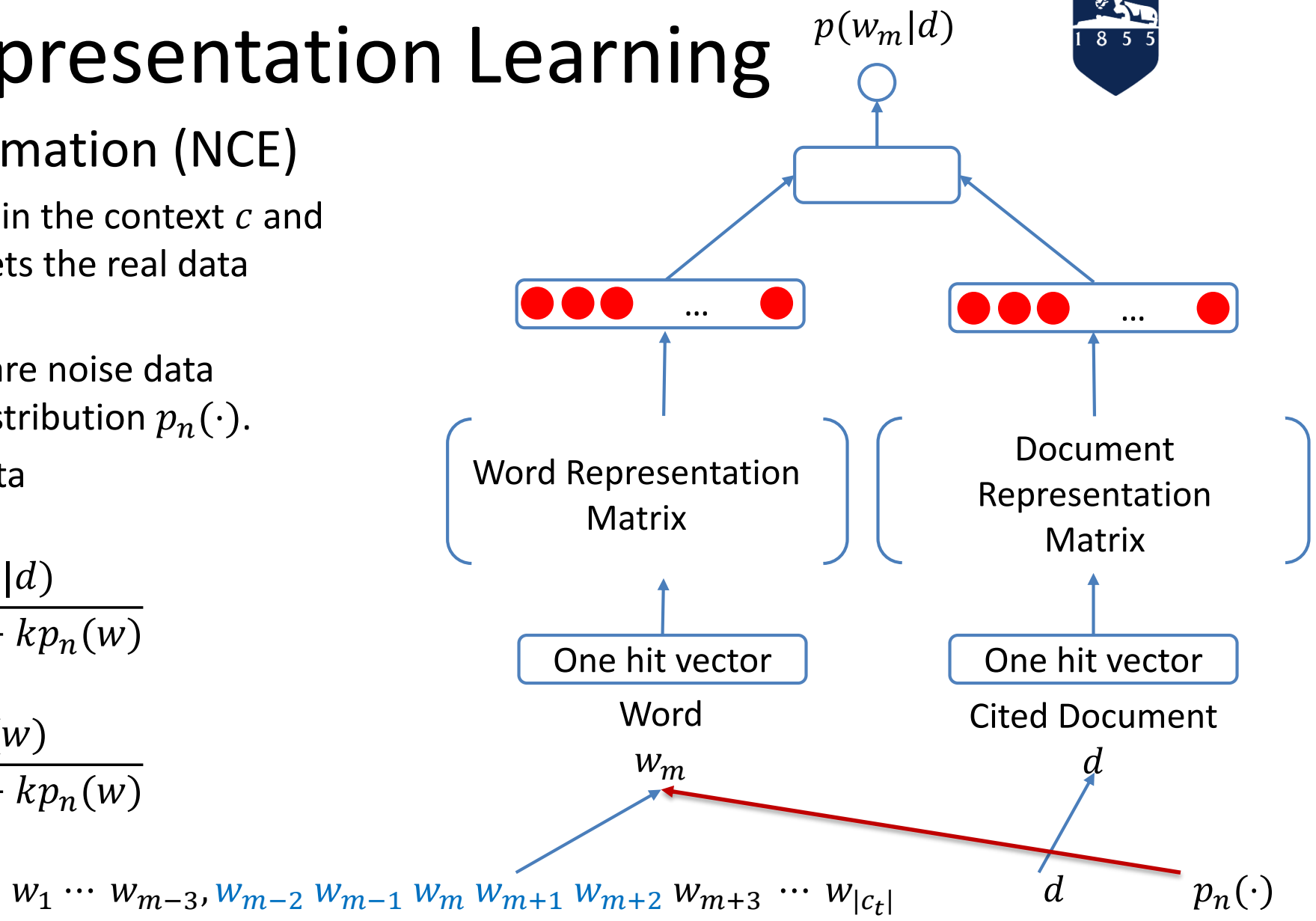


Document Representation Learning

- Noise-contrastive estimation (NCE)
  - Each word w that appears in the context c and the cited document d meets the real data distribution $p_\theta(w|d)$.
  - Any other random words are noise data generated from a noise distribution $p_n(\cdot)$.
  - k times noise than real data

$$p(1|w,d,\theta) = \frac{p_\theta(w|d)}{p_\theta(w|d) + kp_n(w)}$$

$$p(0|w,d,\theta) = \frac{kp_n(w)}{p_\theta(w|d) + kp_n(w)}$$



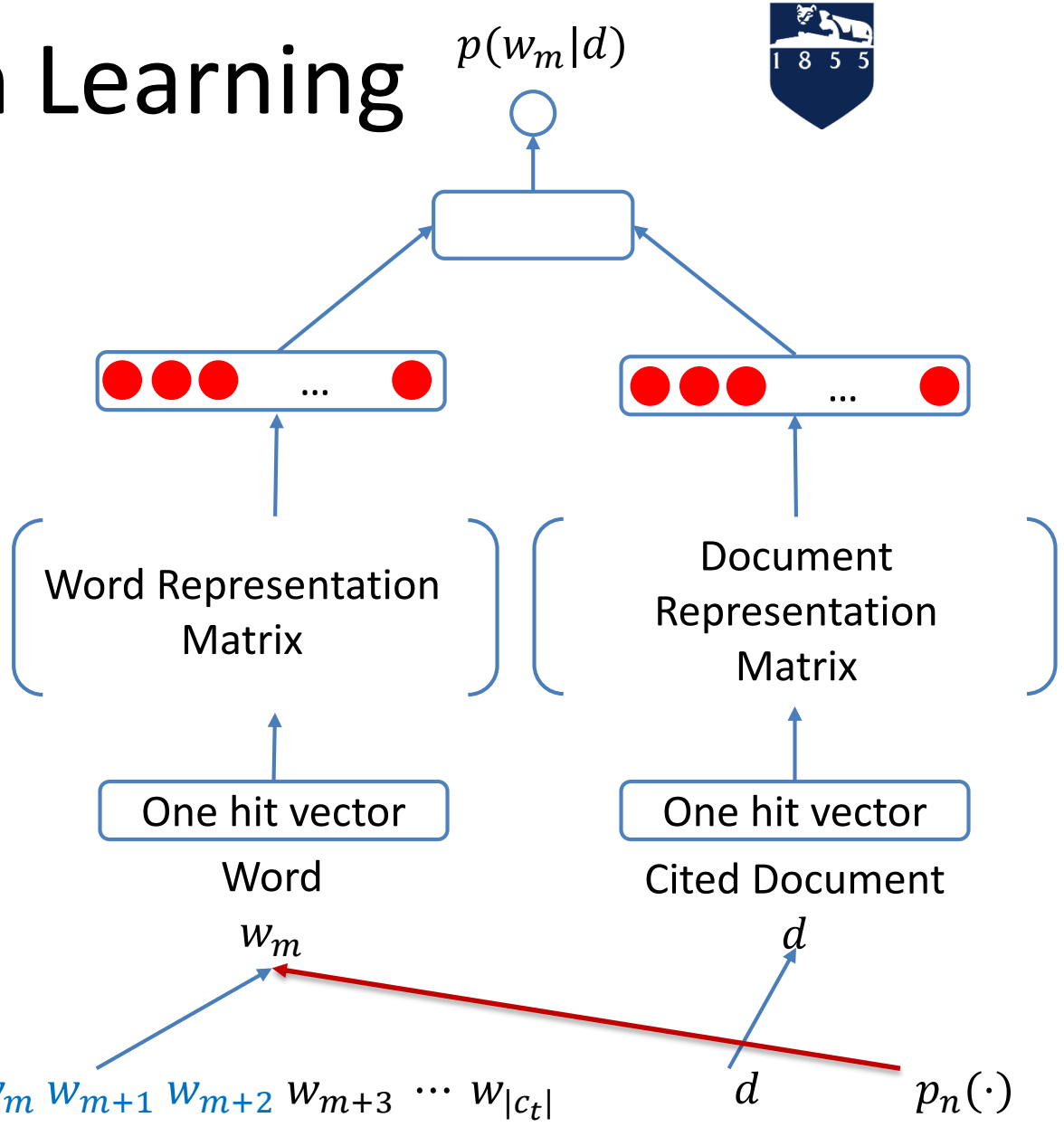
Document Representation Learning

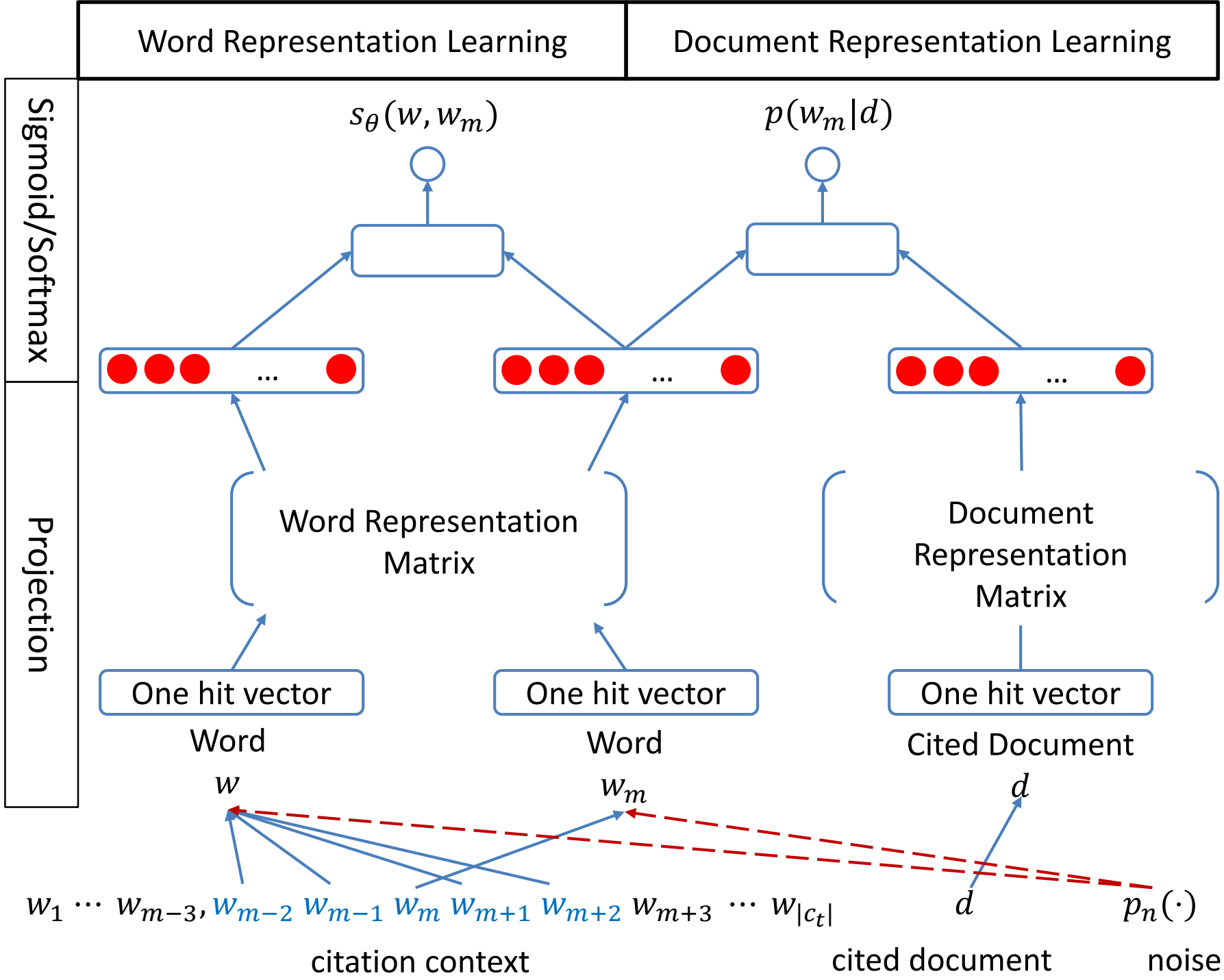
- Noise-contrastive estimation (NCE)
  - The likelihood of binary classification is a Bernoulli distribution.
  - Objective function:

$$\ell_t(\theta) = \log \frac{p_\theta(w|d_t)}{p_\theta(w_{t_i}|d_t) + kp_n(w_{t_i})} + \sum_{i=1}^{k}\left[\log \frac{kp_n(w_{n_i})}{p_\theta(w_{n_i}|d) + kp_n(w_{n_i})}\right]$$

$$p_\theta(w|d) = \exp\big(s_\theta(w,d)\big) \cdot Z(d)$$

The parameters of the neural network are θ and $Z(d)$.




Word Representation Learning
Document Representation Learning
Sigmoid/Softmax
Projection
$s_\theta(w, w_m)$
$p(w_m|d)$
Word Representation Matrix
Document Representation Matrix
One hit vector
One hit vector
One hit vector
Word
Word
Cited Document
w
w_m
d
$w_1 \cdots w_{m-3}, w_{m-2}\ w_{m-1}\ w_m\ w_{m+1}\ w_{m+2}\ w_{m+3} \cdots w_{|c_t|}$
d
$p_n(\cdot)$
citation context
cited document
noise

Citation Recommendation

- Using the fine-tuned word and document representations, we can get the normalized probability distribution $p(w|d)$.

- The table of $p(d|w)$ is pre-calculated using Bayes' rule and stored as an inverted index.

- Given a query $q = [w_1, w_2, \dots, w_{|q|}]$, the task is to recommend a list of document $R = [d_1, d_2, \dots, d_N]$ that need to be cited.

$$p(d_i|q) = \sum_{j=1}^{|q|} p(d_i|w_j) p(w_j|q)$$

where $p(w_j|q)$ can be measured using TF-IDF.

Experiments

- Data:
  - A snapshot of CiteSeer paper and citation database. (Oct. 2013)
    - Training: paper crawled before 2011 (included)
    - Testing: paper crawled after 2011
  - Citations are extracted along with their citation contexts.
    - Training: $|\mathrm{C}| = 8{,}992{,}476$ pairs of citations and citation contexts
    - Testing: $1{,}628{,}698$ pairs.
  - Word vocabulary size: $|V| = 281{,}817$.
  - Cited documents: $|D| = 329{,}365$.

Experiments

- Metrics:
  - Mean average precision(MAP):

$$\text{MAP}=\frac{\sum_{q\in \text{Q}} \text{AveP}(q)}{|Q|} \quad \text{where} \quad \text{AveP}(q) = \frac{\sum_{k=1}^{n} p(k)\cdot \text{rel}(k)}{|\text{correct recommendations}|}$$

  - Recall:

$$\text{Recall} = \frac{|R_g \cap R_r|}{R_g}$$

  - Mean Reciprocal Rank (MRR):

$$\text{MRR} = \frac{1}{|Q|}\sum_{q\in \text{Q}} \frac{1}{\text{rank}_q}$$

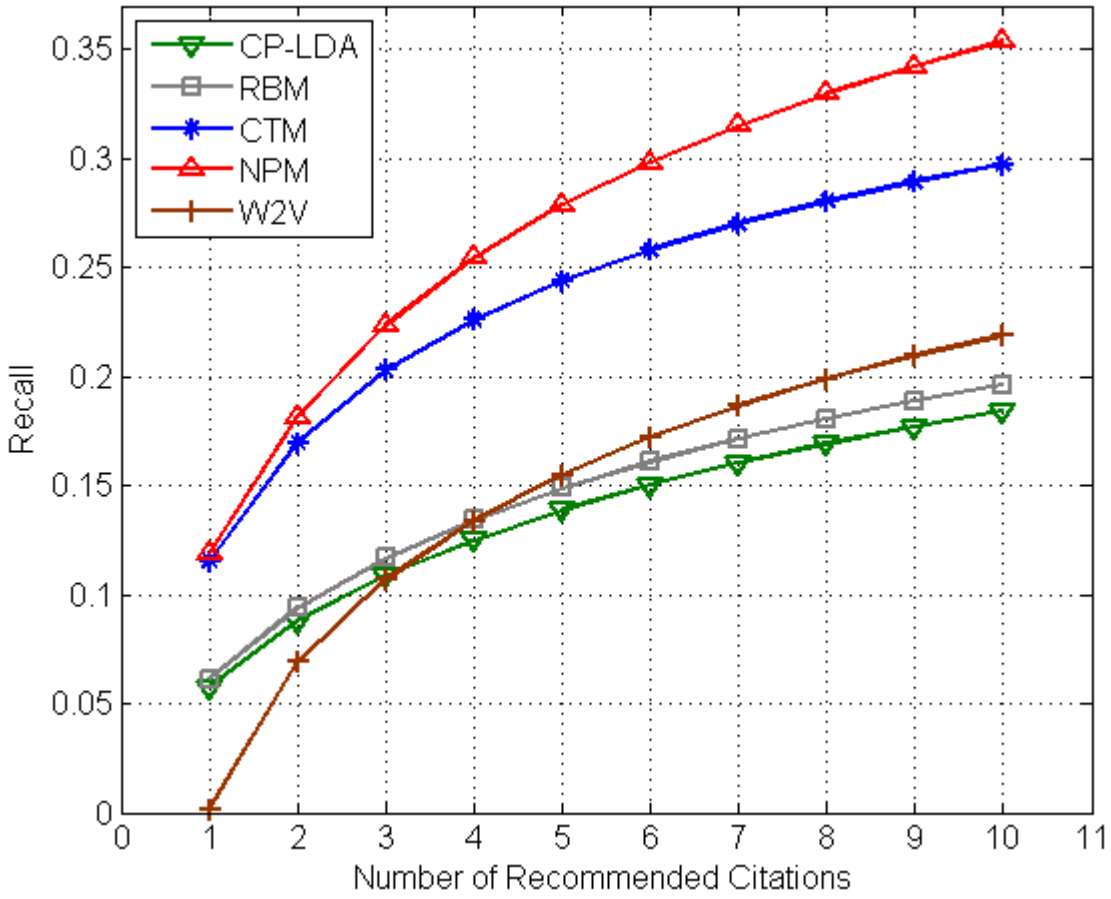
  - Normalized Discounted Cumulative Gain (nDCG):

$$\text{DCG} = rel_1 + \sum_{i=2}^{n} \frac{rel_i}{\log_2 i} \quad \text{and} \quad \text{nDCG}=\frac{\text{DCG}}{\text{IDCG}} \quad \text{where IDCG is idea DCG score}$$

Baselines and Parameter Settings

- Cite-PLSA-LDA (CP-LDA) [Kataria, Mitra, and Bhatia 2010]
  - Number of topic: 600
- Restricted Boltzmann Machine (RBM) [Tang and Zhang 2009]
  - Two layer RBM
  - hidden layer size 600
- Citation Translation Model (CTM) [Huang et al. 2012b]
  - GIZA++ toolkit
  - training iteration: 20
- Word2vec Model(W2V) [Mikolov et al. 2013]
  - Cited documents are regarded as “words” (one document uses a unique token when cited by different papers).
  - Representation vector dimension: 600

Baseline Comparison



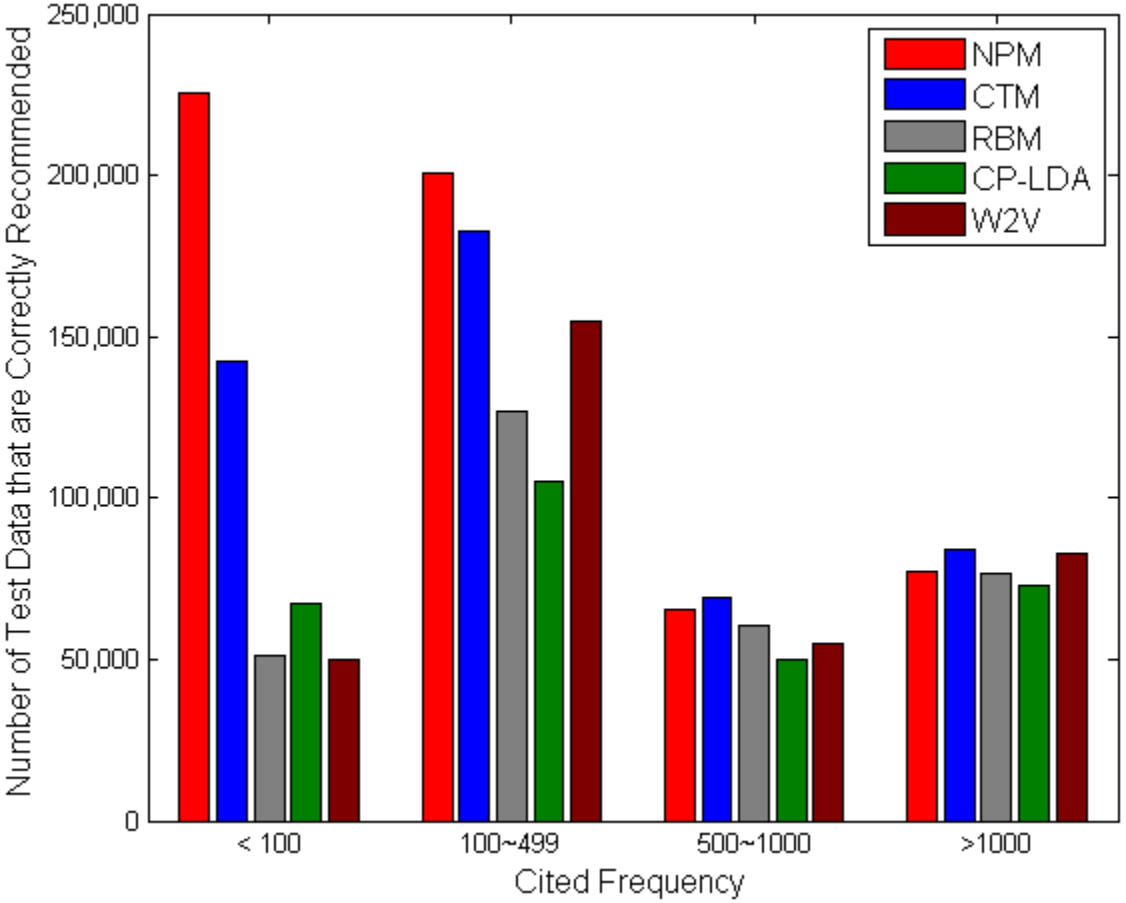
Recall as the number of recommended citations ranges from 1 to 10.

Model	MRR	MAP	nDCG
CP-LDA	0.0916	0.0912	0.1288
RBM	0.0997	0.0982	0.1476
W2V	0.0662	0.0663	0.1356
CTM	0.1687	0.1681	0.2261
NPM	**0.1843***	**0.1835***	**0.2566***

MRR, MAP, nDCG scores for top 10 recommendations; * indicates when NPM better than CTM is statistically significant ($p < 0.001$).

- The proposed model improves the overall recommendation with a 5% gain on Recall@10 and 2% gain on MAP compared to the second best model CTM.
- The proposed model generates better ranked recommendations.
- The significant tests show that the improvements are statistically significant.

Baseline Comparison



Performance versus papers' cited frequency.

- According to the citation counts of the cited documents, we split the test set into four intervals: <100, 100~500, 500~1000, and >1000.
- For each interval, we plot the number of test data that are correctly recommended by each model.
- Our proposed model is particularly good at recommending papers that are not frequently cited (less than 100 citations).
- This result indicates that our model can learn a good representation for documents, even with a small number (less than 100) of training data.

Model Parameters

- Noise Distribution p_n and Number of Samples k

# of Noise k	Uniform Noise	Frequency-based Noise
100	0.3096	0.2913
300	0.3275	0.3165
500	0.3491	0.3297
1,000	**0.3537**	**0.3450**

Recall@10 versus noise distributions and sample size.

- Dimension n

Dimension n	MRR	MAP	nDCG	Recall@10
100	0.1681	0.1692	0.2297	0.3118
300	0.1709	0.1702	0.2372	0.3247
600	**0.1843**	**0.1835**	**0.2566**	**0.3537**

MRR, MAP, nDCG and Recall@10 versus word and document representation dimension.

Recommendation Examples

The query is a citation context extracted from paper ***Fast computation of simrank for static and dynamic information networks***:

There is a lot of research work on static information network analysis, including..., and node ranking[, *]...*

Query and Ground Truth	NPM	CTM	RBM	CP-LDA	W2V
There is a lot of research work on static information network analysis, including ... , and node ranking [1, 2], ... [1] The PageRank citation ranking [2] Authoritative sources in a hyperlinked environment	[O] Authoritative sources in a hyperlinked environment [X] The anatomy of a large-scale hypertextual Web search engine [X] Topic-sensitive PageRank [X] Improved Algorithms for Topic Distillation in Hyperlinked Environments [O] The PageRank citation ranking	[X] Modern Information Retrieval [X] An Efficient Boosting Algorithm for Combining Preferences [X] Optimizing search engines using clickthrough data [X] Learning to rank using gradient descent [X] Chord: A Scalable Peer-to-Peer Lookup Service for Internet Applications	[X] A survey of active network research [X] Statistical mechanics of complex networks [X] How to model an internetwork [X] Network information flow [X] SNORT - lightweight intrusion detection for networks	[X] Network Information Flow [X] Linear network coding [X] Polynomial Time Algorithms for Network Information Flow [X] An algebraic approach to network coding [X] Network Coding for Large Scale Content Distribution	[X] Modern Information Retrieval [X] The anatomy of a large-scale hypertextual Web search engine [X] Chord: A Scalable Peer-to-Peer Lookup Service for Internet Applications [X] A Scalable Content-addressable Network [O] Authoritative sources in a hyperlinked environment

Related Works: Global Recommendation

- Collaborative filtering:
  - McNee, et al. On the recommending of citations for research papers. CSCW '02
    - Recommends citations based on partial list of references
- Text similarity and bibliography similarity:
  - Strohman, et al. Recommending citations for academic papers. SIGIR’07
  - Bethard and Jurafsky. Who should I cite: learning literature search models from citation behavior. CIKM’10
- LDA based model:
  - Kataria, et al. Utilizing context in generative Bayesian models for linked corpus, AAAI’10
    - Recommends topical related documents.

PENNSTATE

Related Works: Local Recommendation

- Restricted Boltzmann Machine:
  - Tang and Zhang. A discriminative approach to topic-based citation recommendation. PAKDD'09
- Context-aware model:
  - He, et al. Context-aware citation recommendation. WWW'10
  - He, et al. Citation recommendation without author supervision. WSDM'11
- Translation model:
  - Lu, et al. Recommending citations with translation model. CIKM'11
  - Huang, et al. Recommending citations: Translating papers into references. CIKM'12

Conclusion

- We propose a neural probabilistic model that learns the probability of citing a paper given a citation context based on distributed representations of words and documents.

- The model was trained and evaluated on the entire CiteSeer dataset which consists of 10,760,318 pairs of citation contexts and cited documents from 1,017,457 papers.

- Compared to other state-of-the-art context-based methods, our model shows significant improvement on various performance metrics, with a 5% gain in recall@10, a 2% gain in MRR and MAP, and a 3% gain in nDCG.

PENNSTATE

1855

THANK YOU!

RefSeer

We don’t keep query and click data!

DEMO: http://refseer.ist.psu.edu/

Q&A

CiteSeer$^{x}_{\beta}$