Recommending Citations: Translating Papers into References

Wenyi Huang[1] Saurabh Kataria[2] Cornelia Caragea[1]
Prasenjit Mitra[1] C. Lee Giles[1] Lior Rokach[3]

[1]Information Sciences & Technology
The Pennsylvania State University

[2]Xerox Research Center Webster

[3]Information Systems Engineering
Ben-Gurion University of the Negev

The Conference on Information and Knowledge Management, 2012

Outline

1. Introduction
   - Problem
   - Motivation
   - Related Work
2. Citation Translation Model
   - Building Up Dictionary
   - Reference Recommendation Using Dictionary
3. Experiment
   - Dataset and Metrics
   - Evaluation

Outline

1. Introduction
   - Problem
   - Motivation
   - Related Work
2. Citation Translation Model
   - Building Up Dictionary
   - Reference Recommendation Using Dictionary
3. Experiment
   - Dataset and Metrics
   - Evaluation

Citation Recommendation

- What is citation recommendation?
- Citation Recommendation:
  - based on a partial list of reference.
  - **based on the content of a manuscript.**

Citation Recommendation

- What is citation recommendation?
- Citation Recommendation:
  - based on a partial list of reference.
  - **based on the content of a manuscript.**

Citation Recommendation

- What is citation recommendation?
- Citation Recommendation:
  - based on a partial list of reference.
  - based on the content of a manuscript.

Citation Recommendation

- What is citation recommendation?
- Citation Recommendation:
  - based on a partial list of reference.
  - **based on the content of a manuscript.**

Outline

1 Introduction
- Problem
- Motivation
- Related Work

2 Citation Translation Model
- Building Up Dictionary
- Reference Recommendation Using Dictionary

3 Experiment
- Dataset and Metrics
- Evaluation

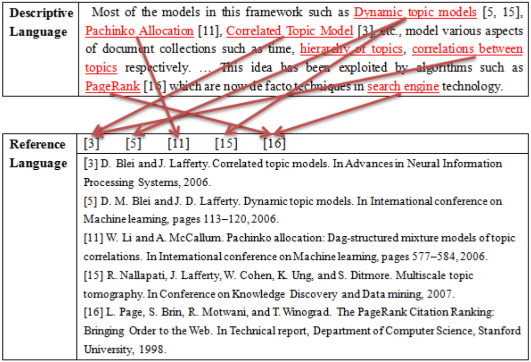
Motivation

A research paper is written using two different “languages”:

- **Descriptive language**, consisting of citation words used in the paper before the reference section;
- **Reference language**, consisting of references, where each referenced paper is considered as a “word”.

Motivation

A citation's context contains explicit words explaining the citation



Descriptive Language	Most of the models in this framework such as Dynamic topic models [5, 15], Pachinko Allocation [11], Correlated Topic Model [3], etc., model various aspects of document collections such as time, hierarchy of topics, correlations between topics respectively. ... This idea has been exploited by algorithms such as PageRank [16] which are now de facto techniques in search engine technology.

Reference Language	[3] [5] [11] [15] [16]
	[3] D. Blei and J. Lafferty. Correlated topic models. In Advances in Neural Information Processing Systems, 2006. [5] D. M. Blei and J. D. Lafferty. Dynamic topic models. In International conference on Machine learning, pages 113–120, 2006. [11] W. Li and A. McCallum. Pachinko allocation: Dag-structured mixture models of topic correlations. In International conference on Machine learning, pages 577–584, 2006. [15] R. Nallapati, J. Lafferty, W. Cohen, K. Ung, and S. Ditmore. Multiscale topic tomography. In Conference on Knowledge Discovery and Data mining, 2007. [16] L. Page, S. Brin, R. Motwani, and T. Winograd. The PageRank Citation Ranking: Bringing Order to the Web. In Technical report, Department of Computer Science, Stanford University, 1998.

Figure : An example of translation from the descriptive language to the reference language

Outline

1. Introduction
   - Problem
   - Motivation
   - Related Work
2. Citation Translation Model
   - Building Up Dictionary
   - Reference Recommendation Using Dictionary
3. Experiment
   - Dataset and Metrics
   - Evaluation

Citation Recommendation

- None Content Based
  - Collaborative filtering (McNee, et al., 2002).
- Content Based
  - Feature Based (Strohman, et al., 2007; Bethard and Jurafsky, 2010)
  - Topic Model based recommendation: cite-PLSA-LDA (Kataria, et al., 2010),
  - Citation Context Based (He, 2010; He, 2011)

Citation Recommendation

- None Content Based
  - Collaborative filtering (McNee, et al., 2002).
- Content Based
  - Feature Based (Strohman, et al., 2007; Bethard and Jurafsky, 2010)
  - Topic Model based recommendation: cite-PLSA-LDA (Kataria, et al., 2010),
  - Citation Context Based (He, 2010; He, 2011)

Citation Recommendation

- None Content Based
  - Collaborative filtering (McNee, et al., 2002).
- Content Based
  - Feature Based (Strohman, et al., 2007; Bethard and Jurafsky, 2010)
  - Topic Model based recommendation: cite-PLSA-LDA (Kataria, et al., 2010),
  - Citation Context Based (He, 2010; He, 2011)

Citation Recommendation

- None Content Based
  - Collaborative filtering (McNee, et al., 2002).
- Content Based
  - Feature Based (Strohman, et al., 2007; Bethard and Jurafsky, 2010)
  - Topic Model based recommendation: cite-PLSA-LDA (Kataria, et al., 2010),
  - Citation Context Based (He, 2010; He, 2011)

Citation Recommendation

- None Content Based
  - Collaborative filtering (McNee, et al., 2002).
- Content Based
  - Feature Based (Strohman, et al., 2007; Bethard and Jurafsky, 2010)
  - Topic Model based recommendation: cite-PLSA-LDA (Kataria, et al., 2010),
  - Citation Context Based (He, 2010; He, 2011)

Citation Recommendation

- None Content Based
  - Collaborative filtering (McNee, et al., 2002).
- Content Based
  - Feature Based (Strohman, et al., 2007; Bethard and Jurafsky, 2010)
  - Topic Model based recommendation: cite-PLSA-LDA (Kataria, et al., 2010),
  - Citation Context Based (He, 2010; He, 2011)

Translation Model

Many tasks in IR and NLP also adopt the translation model to estimate the relationship between two different objects:

- Question answering (Murdock 2004)
- Sentence retrieval (Murdock and Croft, 2005)
- Tag suggestions (Liu, et al., 2011)

Translation Model

Many tasks in IR and NLP also adopt the translation model to estimate the relationship between two different objects:

- Question answering (Murdock 2004)
- Sentence retrieval (Murdock and Croft, 2005)
- Tag suggestions (Liu, et al., 2011)

Translation Model

Many tasks in IR and NLP also adopt the translation model to estimate the relationship between two different objects:

- Question answering (Murdock 2004)
- Sentence retrieval (Murdock and Croft, 2005)
- Tag suggestions (Liu, et al., 2011)

Outline

1. Introduction
   - Problem
   - Motivation
   - Related Work
2. Citation Translation Model
   - Building Up Dictionary
   - Reference Recommendation Using Dictionary
3. Experiment
   - Dataset and Metrics
   - Evaluation

Constructing Parallel Dataset

Suppose a descriptive language with k citation contexts $d = [c_1, \cdots, c_k]$ and the reference language $r = [r_1, \cdots, r_m]$, We construct the parallel data:

Parallel Data

$$\begin{array}{ll} \text{Source} & t_{c_1,1}, \cdots, t_{c_1,|c_1|}, \cdots, t_{c_k,1}, \cdots, t_{c_k,|c_k|} \\ & \Downarrow \\ \text{Target} & r_1, r_2, \cdots, r_m \end{array}$$

where $t_{c_i,j}$ is the jth term appearing in the ith citation context of d and r_i is the ith cited paper in r.

Learning Translation Model

Using IBM Model-1 models, the alignment from source language $d = [t_1, \cdots, t_l]$ to target language $r = [r_1, \cdots, r_m]$ is described by a hidden variable $A = [a_1, \cdots, a_m]$.

Translation Model

$$\begin{array}{ll} \text{Maximize} & \Pr(r|d) = \sum_{a_1=1}^{l} \cdots \sum_{a_m=1}^{l} \prod_{i=1}^{m} \Pr(r_i|t_{a_i}) \\ \text{Subject to} & \sum_{i=1}^{m} \Pr(r_i|t_j) = 1 \qquad j = 1, 2, \cdots, l \end{array}$$

where $\Pr(r_i|t_{a_i})$ is the probability of citing r_i given a term t_{a_i}.

Outline

1. Introduction
   - Problem
   - Motivation
   - Related Work
2. Citation Translation Model
   - Building Up Dictionary
   - Reference Recommendation Using Dictionary
3. Experiment
   - Dataset and Metrics
   - Evaluation

Reference Recommendation Using Dictionary

- Translation table between two vocabularies in the form of triplet entries$\langle t_i, r_j, Pr(r_j|t_i)\rangle$
- Given Query $Q = [t_1, \cdots, t_l]$, the task is to recommend a list of references $R = [r_1, \cdots, r_m]$.

Ranking Function

$$\Pr(r_i|Q) = \sum_{j=1}^{l} \Pr(r_i|t_j)\Pr(t_j|Q)$$

Reference Recommendation Using Dictionary

- Translation table between two vocabularies in the form of triplet entries$\langle t_i, r_j, Pr(r_j|t_i)\rangle$
- Given Query $Q = [t_1, \cdots, t_l]$, the task is to recommend a list of references $R = [r_1, \cdots, r_m]$.

Ranking Function

$$\Pr(r_i|Q) = \sum_{j=1}^{l} \Pr(r_i|t_j)\Pr(t_j|Q)$$

Reference Recommendation Using Dictionary

- Translation table between two vocabularies in the form of triplet entries$\langle t_i, r_j, Pr(r_j|t_i)\rangle$
- Given Query $Q = [t_1, \cdots, t_l]$, the task is to recommend a list of references $R = [r_1, \cdots, r_m]$.

Ranking Function

$$\Pr(r_i|Q) = \sum_{j=1}^{l} \Pr(r_i|t_j)\Pr(t_j|Q)$$

Outline

1. Introduction
   - Problem
   - Motivation
   - Related Work
2. Citation Translation Model
   - Building Up Dictionary
   - Reference Recommendation Using Dictionary
3. Experiment
   - Dataset and Metrics
   - Evaluation

Dataset

CiteSeer used used for citation recommendation by Kataria, et al (2010), Tang and Zhang (2009).

CiteULike from November 2005 to January 2008.

Data	D	C	W_C	R	$\bar{N}_c$
CiteSeer	$3,312$	$26,597$	$21,982$	$2,138$	18.01
CiteULike	$14,418$	$40,720$	$52,631$	$5,484$	8.61

Table : D is the number of documents, C is the number of citation contexts, W_C is the number of unique words in citation contexts, R is the number of unique references, and $\bar{N}_c$ is the number of average citations a paper has.

Metrics

Precision, Recall, F-measure

$$p. = \frac{|R_g \cap R_r|}{R_r}, r. = \frac{|R_g \cap R_r|}{R_g}, f. = \frac{2p. \times r.}{p. + r.}$$

Binary Preference Measure (Bpref)

$$\text{Bpref} = \frac{1}{|R|} \sum_{r \in R} 1 - \frac{|i \text{ ranked higher than } r|}{|S|}$$

Mean Reciprocal Rank (MRR)

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

Outline

1. Introduction
   - Problem
   - Motivation
   - Related Work
2. Citation Translation Model
   - Building Up Dictionary
   - Reference Recommendation Using Dictionary
3. Experiment
   - Dataset and Metrics
   - Evaluation

Runtime and Comparing Results

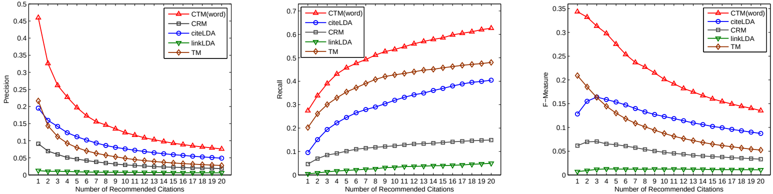
	Training		Recommending	
	CiteSeer	CiteULike	CiteSeer	CiteULike
link-LDA	622.490s	20824.61s	1.790s	34.865s
CRM	-	-	2006.032s	3012.003s
cite-LDA	594.115s	8949.210s	1.845s	20.154s
TM	573.891s	866.227s	6287.421s	9972.11s
CTM	**53.372s**	**71.460**s	**1.480s**	**4.904s**

Table : Runtime on `CiteSeer` and `CiteULike` dataset.

	CiteSeer		CiteULike	
	Bpref	MRR	Bpref	MRR
link-LDA	0.064	0.028	0.027	0.013
CRM	0.097	0.238	0.054	0.072
cite-LDA	0.459	0.285	0.260	0.143
TM	0.422	0.288	0.393	0.285
CTM	**0.645**	**0.529**	**0.627**	**0.467**

Table : Bpref and MRR metrics on `CiteSeer` and `CiteULike` dataset with 20 recommended paper.

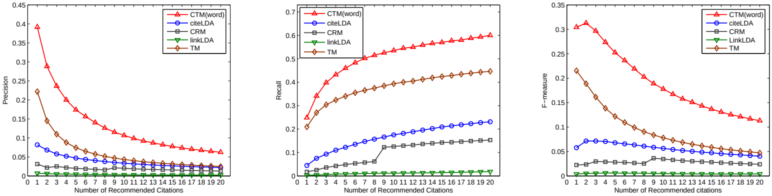
Comparing Results



(a) Precision (b) Recall (c) F-measure

Figure : Precision, recall and F-measure of different methods on `CiteSeer` dataset with recommended citations range from 1 to 20.

Comparing Results



(a) Precision

(b) Recall

(c) F-measure

Figure : Precision, recall and F-measure of different methods on `CiteULike` dataset with recommended citations range from 1 to 20.

Contribution

- We propose to represent the cited papers by unique IDs,regarding them as “words” in a novel language, and then use translation model to estimate the translation probability of a ID given citing words.
- CTM increase the precision, recall and f-measure by at least 5% to 10%, respectively, compared with the state-of-the-art approaches.
- On large datasets, CTM runs at least 100 times faster in the training stage and 5 to 600 times faster in the recommending stage.

Contribution

- We propose to represent the cited papers by unique IDs,regarding them as “words” in a novel language, and then use translation model to estimate the translation probability of a ID given citing words.
- CTM increase the precision, recall and f-measure by at least 5% to 10%, respectively, compared with the state-of-the-art approaches.
- On large datasets, CTM runs at least 100 times faster in the training stage and 5 to 600 times faster in the recommending stage.

Contribution

- We propose to represent the cited papers by unique IDs,regarding them as “words” in a novel language, and then use translation model to estimate the translation probability of a ID given citing words.
- CTM increase the precision, recall and f-measure by at least 5% to 10%, respectively, compared with the state-of-the-art approaches.
- On large datasets, CTM runs at least 100 times faster in the training stage and 5 to 600 times faster in the recommending stage.

Q&A

Thank you!

RefSeer

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