RefSeer: A Citation Recommendation System

Wenyi Huang
harrywy@gmail.com

Zhaohui Wu
laowuz@gmail.com

Prasenjit Mitra
pmitra@ist.psu.edu

C. Lee Giles
giles@ist.psu.edu

Information Sciences and Technology
The Pennsylvania State University
University Park, PA 16802

Outline

- **Introduction**
  - Problem Definition
  - Existing Systems
- **The New RefSeer**
  - Metadata
  - Global Recommendation
  - Local Recommendation
  - Complexity Analysis
- **Evaluation**
  - Dataset and Metrics
  - Recommendation Results

Problem Definition

- What is automatic citation recommendation?
  - Assumption – have a document we want citations for

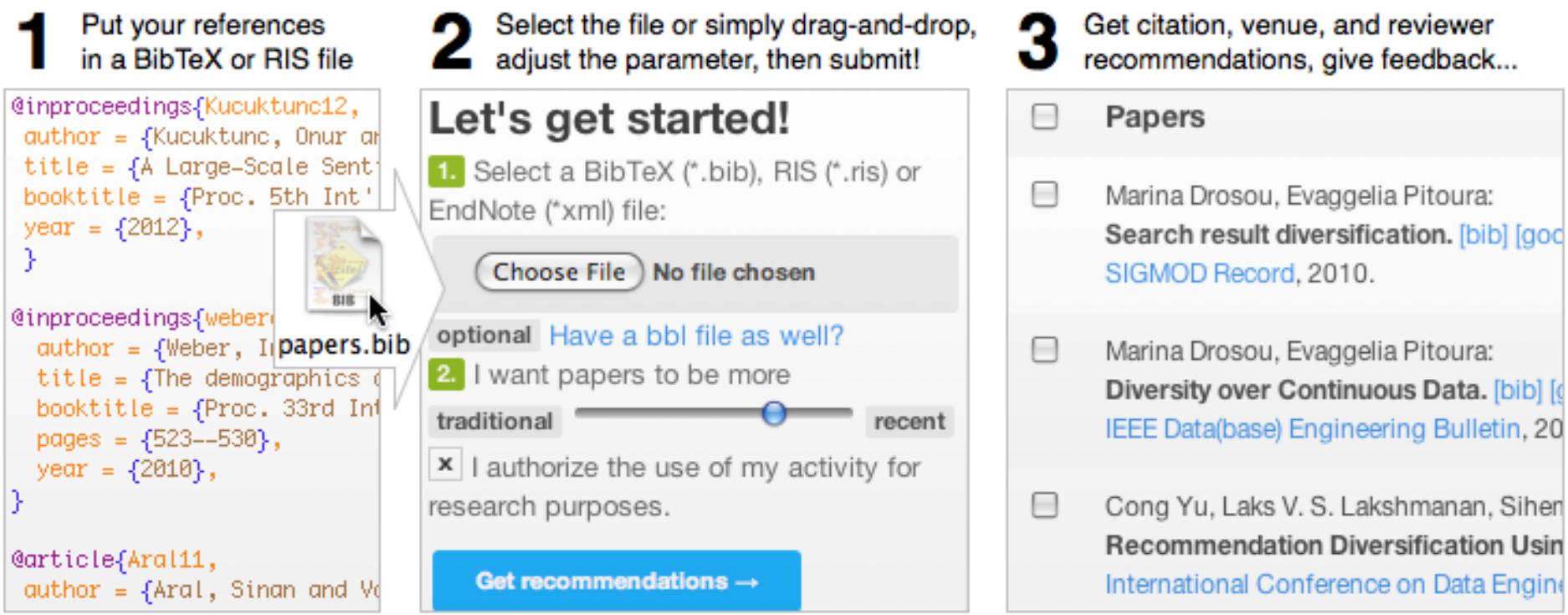
- Recommend Citations for a given document:
  - based on a partial list of citations in the document
    - Similar citations in other documents.
  - **based on the content in the document**
    - **Similar documents**
      - **Topic based**
      - **Citation context based**

Existing Systems

- Collaborative filtering (McNee, et al., 2002).
  - based on a partial list of references.
- Feature Based (Strohman, et al., 2007; Bethard and Jurafsky, 2010)
  - Citation network, citation count, recency, topic similarity …
- Restricted Boltzmann Machine (Tang and Zhang, 2009)
  - Topic similarity
- Citation Context Based (He, et al. 2010; 2011)
  - Refseer Prototype
- TheAdvisor (http://theadvisor.osu.edu/)
  - based on a partial list of references. (Kucuktunc, et al. 2012)
- Refseer (http://refseer.ist.psu.edu/)
  - based on the content in the document
    - Topic based (Kataria, et al., 2010)
    - Citation context based (Huang, et al. 2012)

TheAdvisor

- Input: a partial list of references.



O. Kucuktunc et al., “Fast recommendation on bibliographic networks,” in Proc. of ASONAM2012. IEEE Computer Society, 2012, pp. 480–487.

Refseer Prototype

- Input: the content of a manuscript
- Recommendation model (He, et al 2010; 2011).
  - Citation context similarity measure
- Built on a cluster of 8 nodes with each node has 8 2.57GHz CPUs and 32GB memory.
- The complexity of generating the candidate list is linear to the size of CiteSeer repository.
- The computation of ranking score and reranking are also time-consuming.
- Extremely slow! Not efficient for public services!

Q. He et al., “Context-aware citation recommendation,” in Proc. of the 19th Int. Conf. on World wide web. New York, NY, USA: ACM, 2010, pp. 421–430.

The New RefSeer

- Screenshot (http://refseer.ist.psu.edu/)

RefSeer

Recommend

1 - 10 of 100 Results in each topics

Related Topics: Topic 1 Topic 2 Topic 3 Topic 4

topic latent topics model models dirichlet lda words semantic document mixture plsa multinomial allocation blei distribution text generative word probabilistic

Latent dirichlet allocation

by David M. Blei, Andrew Y. Ng, Michael I. Jordan, John Lafferty, *Journal of Machine Learning Research , 2003* .

Abstract - Cited by 1277 (44 self) - Like

Probabilistic Latent Semantic Indexing

by Thomas Hofmann , , *1999* .

Abstract - Cited by 521 (7 self) - Like

Probabilistic Latent Semantic Analysis

by Thomas Hofmann, *In Proc. of Uncertainty in Artificial Intelligence, UAI'99*

Abstract - Cited by 362 (5 self) - Like

Most of the models in this framework such as Dynamic topic models, Pachinko Allocation, Correlated Topic Model, etc., model various aspects of document collections such as time, hierarchy of topics, correlations between topics respectively. [1]

However, all the above mentioned models ignore a rich feature that contains valuable information, namely, the citation or hyperlink structure. [2]

It is a known fact in information retrieval that a citation between two documents not only indicates topical similarity of the two documents but also authoritativeness of the cited document. [3]

This idea has been exploited by algorithms such as PageRank which are now de facto techniques in search engine technology. [4]

1 - 10 of 150 Results

Sentence No: 1 2 3 4

Pachinko allocation: DAG-structured mixture models of topic correlations

by Wei Li, Andrew Mccallum, *In Proceedings of the 23rd International Conference on Machine Learning , 2006* .

Abstract - Cited by 72 (5 self) - Like

Latent dirichlet allocation

by David M. Blei, Andrew Y. Ng, Michael I. Jordan, John Lafferty, *Journal of Machine Learning Research , 2003* .

Abstract - Cited by 1277 (44 self) - Like

Dynamic topic models

by David M. Blei, John D. Lafferty, *In ICML , 2006* .

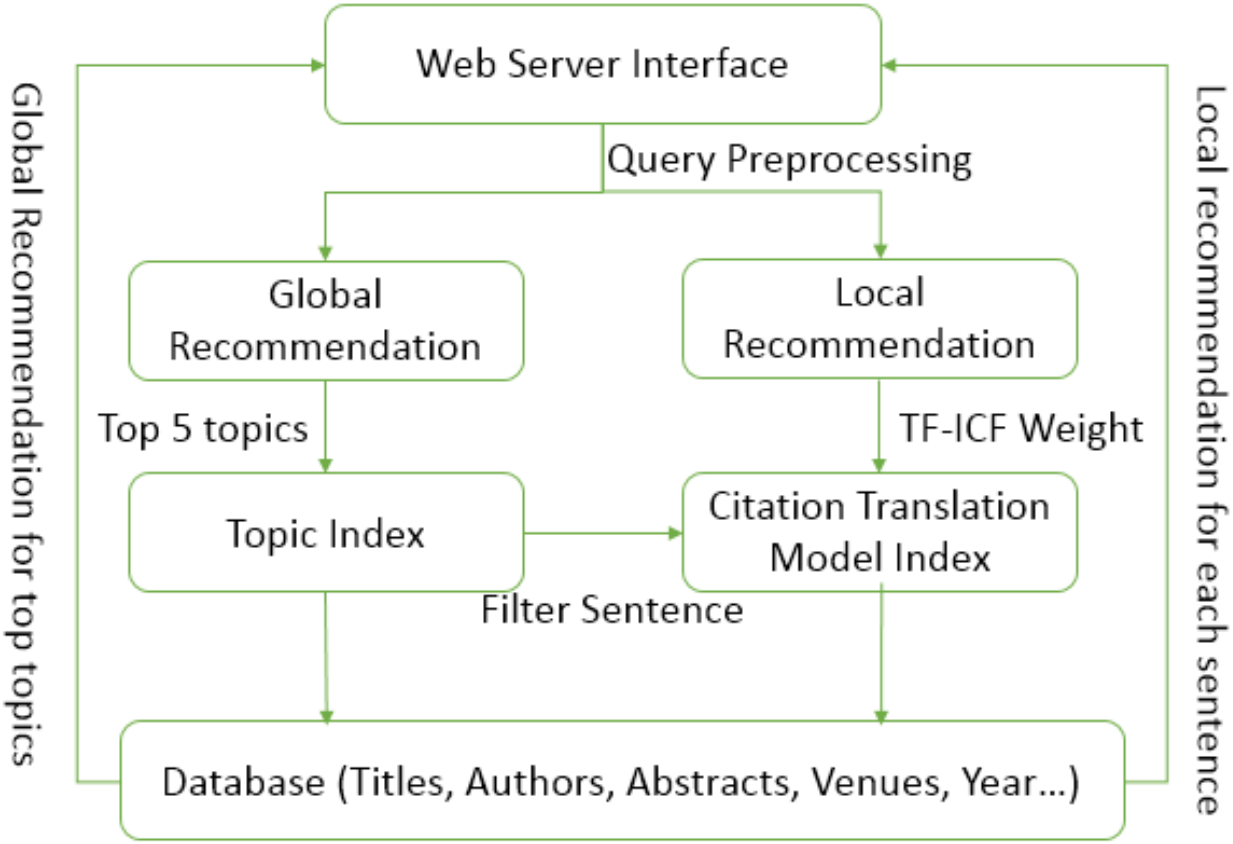
Abstract - Cited by 231 (15 self) - Like

Hierarchically Classifying Documents Using Very Few Words

by Daphne Koller, Mehran Sahami, , *1997* .

Refseer Approach

- Global Recommendation:
  - Recommend topical related documents with respect to the whole input query.
- Local Recommendation:
  - Recommend for document for each sentence that needs citations.



Metadata used (and shared)

- RefSeer uses all paper metadata provided by CiteSeer.
  - 1,017,457 papers as of Oct. 2013.
- Papers' content are parsed and used for training topic based model.
- Citations and citation contexts are extracted for training local recommendation model.
  - 10,760,318 citation relations
  - 83,598,304 citation contexts

Global Recommendation

- Global Recommendation:
  - Recommend topical related documents with respect to the whole input query.
- Training:
  - RefSeer internally computes topical compositions for each paper using Cite-PLSA-LDA mode proposed by S. Kataria et al.
  - Word-topic Pr(t|w) and topic-citation Pr(d|t) distributions were inferred over all documents in CiteSeer repository.
  - Number of topics: 1000.
- Recommendation:
  - For a new query, RefSeer will infer the top 5 topics (at most) using the word-topic distribution.
  - For each top topic, a list of citations will be recommended using topic-citation distribution.

S. Kataria et al., "Utilizing context in generative Bayesian models for linked corpus," in Proc. of the 24th AAAI Conf. on Artificial Intelligence, pp. 1340–1345.

Local Recommendation

- Local Recommendation:
  - Recommend for document for each sentence that needs citations.
- Training:
  - RefSeer uses Citation Translation Model to learn the "translation" probability of citing a document given a word Pr(d|w).
  - We modify the GIZA++ toolkit to learn translation probabilities using IBM Model-1.
- Recommendation:
  - Ranking function:

$$\Pr(d|Q) = \sum_{j=1}^{l} \Pr(d|w_j) \Pr(w_j|Q)$$

  - We use term-frequency-inverse-context-frequency (TF-ICF) to measure Pr(w|Q)
  - Filtering: If the topic distribution of top recommended paper for a sentence has a very low similarity with the topic distribution of the query, we filter out the sentence.

W. Huang et al., "Recommending citations: Translating papers into references," in Proc. of CIKM'12, 2012, pp. 1910–1914.

Complexity Analysis

- Training
  - Global: $O(IKD\overline{N_w})$
  - Local: $O(ID\overline{N_{cc}} \cdot \overline{N_c}^2)$
- Recommending
  - Global: $O(KN_q)$
  - Local: $O(N_q R_q)$

- I : training iterations.
- $\overline{N_w}$: number of words in each paper. $(\overline{N_w} > \overline{N_{cc}} \cdot \overline{N_c})$
- $\overline{N_{cc}}$: number of words in citation context. (200)
- $\overline{N_c}$: number of citations each paper has. (around 20)
- K : number of topics. (200 ~ 500)
- D : size of corpus.
- N_q : number of words in query.
- R_q : number of dictionary entries for each word. (20 ~ 50)

	Training		Recommending	
	Global	Local	Global	Local
CiteSeer(small)	594.115s	53.372s	1.845s	1.480s
CiteULike	8949.210s	71.460s	20.154s	4.904s
CiteSeer(whole)	4d01h49m	5h32m03s	<5ms	<5ms

Table 2: Training and recommending time cost on CiteSeer(small), CiteULike dataset and whole CiteSeer database. The recommending time for CiteSeer(whole) is per query.

Dataset and Metrics

- Dataset:

Data	D	C	R	$\overline{N_c}$
CiteSeer(small)	3,312	26,597	2,138	18.01
CiteULike	14,418	40,720	5,484	8.61
CiteSeer(whole)	1,017,457	10,760,318	856,758	20.73

Table 1: D is the number of documents, C is the number of citation contexts, R is the number of unique references, and $\overline{N_c}$ is the number of average citations a paper has.

- Metrics:
  - Precision, Recall
  - Bpref (Binary Preference Measure):
    - the inversed fraction of irrelevant documents that are retrieved before relevant ones.
  - MRR (Mean Reciprocal Rank):
    - the harmonic mean of the ranks.

Recommendation Results

	CiteSeer(small)		CiteULike	
	Bpref	MRR	Bpref	MRR
Global	0.459	0.285	0.260	0.143
Local	0.645	0.529	0.627	0.467

Table 3: Bpref and MRR on CiteSeer(small) and CiteULike dataset with 20 recommended paper.

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

$$MMR = \frac{1}{|Q|}\sum_{q \in \mathrm{Q}} \frac{1}{\mathrm{rank}_q}$$

- The first correct recommendation will most likely appear among the top 2 for local recommendation.
- The first correct recommendation mostly appears around top 5 recommendations for global recommendation.

Recommendation Results

	CiteSeer(small)		CiteULike	
	Precision	Recall	Precision	Recall
Global	0.13	0.38	0.08	0.49
Local	0.15	0.48	0.12	0.57

Table 4: Precision and Recall on CiteSeer(small) and CiteULike dataset with 10 recommended paper.

- Within top 10 recommendations, local recommendation achieves a recall score around 50%.
- There will be at least 1 correct recommendation among top 10 global recommendations.

One more Example

- Title: Adaptive Methods for the Computation of PageRank
- Abstract (Input as query):

We observe that the convergence patterns of pages in the PageRank algorithm have a nonuniform distribution. Specifically, many pages converge to their true PageRank quickly, while relatively few pages take a much longer time to converge. Furthermore, we observe that these slow-converging pages are generally those pages with high PageRank. We use this observation to devise a simple algorithm to speed up the computation of PageRank, in which the PageRank of pages that have converged are not recomputed at each iteration after convergence. This algorithm, which we call Adaptive PageRank, speeds up the computation of PageRank by nearly 30%

Example

Topic Based Recommendation

1. PageRank computation and the structure of the web: Experiments and algorithms.
2. Improved algorithms for topic distillation in a hyperlinked environment.
3. Automatic resource compilation by analyzing hyperlink structure and associated text
4. A new approach to topic specific web resource discovery.
5. Matrix Computations
6. Probability and Random Processes
7. Efficient computation of PageRank
8. Topic-sensitive PageRank
9. The second eigenvalue of the Google matrix
10. Scaling personalized web search
11. The PageRank citation ranking: Bringing order to the web.
12. Authoritative sources in a hyperlinked environment

Context Based Recommendation

1. The PageRank Citation Ranking: Bringing Order to the Web
2. The Anatomy of a Large-Scale Hypertextual Web Search Engine
3. Authoritative Sources in a Hyperlinked Environment
4. Improved Algorithms for Topic Distillation in a Hyperlinked Environment
5. Optimizing Search Engines using Clickthrough Data
6. Less is more: Active learning with support vector machines
7. Scaling Personalized Web Search
8. Detecting Intrusions Using System Calls: Alternative Data Models
9. Eddies: Continuously Adaptive Query Processing
10. Rule Discovery From Time Series
11. Summarizing Text Documents: Sentence Selection and Evaluation Metrics
12. Hierarchical Reinforcement Learning with the MAXQ Value Function Decomposition

Conclusion

- RefSeer: a citation recommendation system which recommend citations based on the content of paper manuscript.

- RefSeer is very efficient and scalable. (both training and recommending)

- Future directions:
  - Automatic citation context generation
  - Automatic related work section generation
  - …

Q&A

- Thank you!
- Try out RefSeer
  - http://refseer.ist.psu.edu/