A Collective Topic model for Milestone Paper Discovery

Ziyu Lu
The University of Hong Kong
Pokfulam Road, Hong Kong
zylu@cs.hku.hk

Nikos Mamoulis
The University of Hong Kong
Pokfulam Road, Hong Kong
nikos@cs.hku.hk

David W. Cheung
The University of Hong Kong
Pokfulam Road, Hong Kong
dcheung@cs.hku.hk

ABSTRACT

Prior arts stay at the foundation for future work in academic research. However the increasingly large amount of publications make it difficult for researchers to effectively discover the most important previous works to the topic of their research. In this paper, we study the automatic discovery of the core papers for a research area. We propose a collective topic model on three types of objects: papers, authors and published venues. We model any of these objects as bags of citations. Based on Probabilistic latent semantic analysis (PLSA), authorship, published venues and citation relations are used for quantifying paper importance. Our method discusses milestone paper discovery in different cases of input objects. Experiments on the ACL Anthology Network (ANN) indicate that our model is superior in milestone paper discovery when compared to a previous model which considers only papers.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Information Search and Retrieval—*clustering, retrieval models*

Keywords

Topic Model; Milestone Paper; Paper Importance

1. INTRODUCTION

Academic literature surveying plays a vital role in academic research; researchers can learn what has been done, what research gaps might exist and what potential research directions to work on. Academic search engines such as Google Scholar [1] and CiteSeerX [2] enable researchers to find related literatures or prior arts. However, the overwhelming number of publications makes it difficult to quickly obtain the most important set of previous work of a subject. Some citation recommendation systems have been designed to recommend appropriate citations for academic works [5, 1]. However, academic search engines and recommendation systems might return qualified sets of papers based on semantic similarity; paper importance has not been considered. It is essential to have some models for *milestone paper discovery*; i.e., discover the core set of papers which can best represent previous works for a research topic. Wang et al. [8] studied topic milestone paper discovery and developed a generative model for theme and topic evolution. They used the idea of modeling each paper as a "bag of citations" and measured paper impacts based on co-citation relations. However, [8] only considered co-citation relations for topic milestone paper discovery. Therefore, their model could be biased against recently published papers (which are rarely cited by others). Thus, the issue of milestone paper discovery is not completely addressed.

In this paper, we propose a collective topic model for objects of different types (i.e., papers, authors, and venues) in academic networks. The collective topic model quantifies paper importance based on authorship, published venue reputation and co-citation relationships in the academic collection. We investigate the identification of milestone papers in different cases. Our experimental results show that paper importance is well captured by our model; authorship and published venues have considerable influence on milestone paper discovery.

[1]http://scholar.google.com
[2]http://citeseerx.ist.psu.edu/index

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
SIGIR'14, July 6–11, 2014, Gold Coast, Queensland, Australia.
Copyright 2014 ACM 978-1-4503-2257-7/14/07 ...$15.00.
http://dx.doi.org/10.1145/2600428.2609499.

2. RELATED WORK

Previous work exists in citation recommendation [5, 1]; i.e., recommending appropriate references to researchers. For example, [5] designed a translation model between citation contexts and reference words, and recommended a list of citations by using long queries such as sentences or a manuscript. Bethard and Jurafsky [1] designed a feature-based learning model for literature retrieval. A list of references are recommended using the abstract of the input object (i.e., a paper) as query. These works perform recommendations based on semantic analysis, but paper importance or ranking is not considered. In addition, some works in topic evolution may enable researchers to see how the research in a particular area evolves. Mei and Zhai [6] used temporal text mining techniques to discover latent themes from text and constructed theme evolution graphs. However, they cannot identify the "micro-view" of a research field, e.g., milestone papers for that area.

Wang et al. [8] used milestone paper discovery as an application when developing a generative topic model for re-

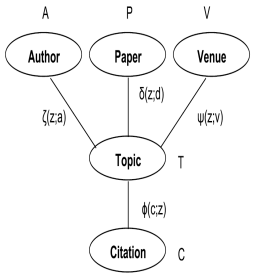

Figure 1: Overview of our collective topic model

search theme evolution. They modeled a paper as a *bag of citations* and use the co-citation relationships for evaluating paper impact between topics. However, their results can be biased against recently published papers, since they did not take into account additional factors that influence the importance of papers, such as authorship and published venues. In our work, we propose a topic model that considers these additional influence factors.

3. PROBABILISTIC TOPIC MODEL

The importance of a paper depends on a variety of factors, including the authority of authors, the publication venue and co-citation relationship with other papers. We borrow the idea of modeling a paper as a *bag of citations* from [8] in order to consider the co-citation factor in the importance of a paper; thus each paper is represented as a bag of citation IDs. Since authors and venues are linked with documents in the academic document collection, we build a "virtual document" for each author and venue by aggregating all documents associated with that author or venue (we call the result *author document* and *venue document*, respectively). This way, for each author or venue we also derive a bag of citation IDs. Based on [3], we assume that the multiple-typed documents (paper document, author document and venue document) have a common set of latent topics and each topic is represented as the distribution over citations. We model these documents using a probabilistic topic model which quantifies the probability of a paper to be cited as paper importance. Then, the problem about milestone paper discovery is defined as follows.

Milestone paper discovery: Given an academic document collection D (papers Q, author set A , published venues V and citations C are known), the model outputs a set of core papers M that includes citationIDs (paperID) ranking based on co-citations at top within a topic or "a document" (The document type might be a paper document, an author document or a venue document).

3.1 Model Description

An overview of our model is shown in Figure 1. Table 1 describes meanings of the notations used in our model. We assume that for all documents there is a common set of k latent topics. Each document is represented as the distribution over topics and each topic is represented as the distribution over citations. Then, the process of generating an academic document is as follows: for each citation in that document, firstly sample a topic z_k according to the distribution from paper_topic distribution $\delta(z;d)$ or author_topic distribution $\zeta(z;a)$ or venue_topic distribution $\psi(z;v)$ based on the document type. Then, draw a citation c from the sampled topic distribution $\phi(:;z_k)$ in topic_citation distribution $\phi(c;z)$.

Table 1: Notations used in our collective model

Symbols	Description
d, a	d for a paper, a for an author
v, c, z	v for a venue, c for a citation, z for a topic
T, k	T for topic set, k for topic number
N	The citation-paper Matrix
U	The citation-author Matrix
E	The citation-venue Matrix
$\phi(c;z)$	The topic-citation distribution
$\delta(z;d)$	The paper-topic distribution
$\zeta(z;a)$	The author-topic distribution
$\psi(z;v)$	The venue-topic distribution
α, β, γ	relative weights for d, a, v

We developed our model based on PLSA [4]. We can have the following joint model for citations based on documents in different types:

$$p(c_i|d_j) = \sum_k \phi(c_i;z_k)\delta(z_k;d_j) \quad (1)$$

$$p(c_i|a_n) = \sum_k \phi(c_i;z_k)\zeta(z_k;a_n) \quad (2)$$

$$p(c_i|v_m) = \sum_k \phi(c_i;z_k)\psi(z_k;v_m) \quad (3)$$

and the parameter set θ to be estimated is:

$$\theta = \{\phi(c|z), \delta(z;d), \zeta(z;a), \psi(z;v) \mid c \in C, d \in D, a \in A, v \in V, z \in T\}$$

In order to estimate the parameters θ, we should maximize the likelihood of the document collection D given θ. The loglikelihood function is represented as

$$L(\theta) = \sum_i (\alpha \sum_j N_{ij} \log p(c_i|d_j) + \beta \sum_n U_{in} \log p(c_i|a_n) + \gamma \sum_m E_{im} \log p(c_i|a_m))$$

N_{ij} indicates the occurrences of citation c_i in paper d_j in the citation-paper matrix, U_{in} the occurrences of citation c_i cited by a_n and E_{in} the occurrences of citation c_i cited by papers in venue v_m. α, β, γ indicate relative weights for three-typed documents d, a, v.

3.2 Parameter Inference

We use the Expectation-Maximization (EM) algorithm for parameter inference. EM iteratively executes two steps, an E-step and a M-step, until $L(\theta)$ converges [2].

Each E-step computes the lower bound function Q of $L(\theta)$. In this process, the posterior probabilities $p(z_k|c,o)$ (o can be d, a, v) are re-computed using the new parameter values from the previous M-step:

$$p(z_k|c_i,d_j) = \frac{\phi(c_i|z_k)\delta(z_k|d_j)}{\sum_k \phi(c_i;z_k)\delta(z_k;d_j)}$$

$$p(z_k|c_i,a_n) = \frac{\phi(c_i;z_k)\zeta(z_k;a_n)}{\sum_k \phi(c_i;z_k)\zeta(z_k;a_n)}$$

Table 2: Topic milestone papers (top-10 papers) for *Sentiment Analysis* from [8].

$\phi(c_i; z_k)$	Venue	Paper Title
0.0785	EMNLP'02	Thumbs Up? Sentiment Classification Using Machine Learning Techniques
0.0672	ACL'02	Thumbs Up Or Thumbs Down? Semantic Orientation Applied To Unsupervised Classification Of Reviews
0.0483	HLT'05	Recognizing Contextual Polarity In Phrase-Level Sentiment Analysis
0.0436	ACL'04	A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based On Minimum Cuts
0.0365	ACL'97	Predicting The Semantic Orientation Of Adjectives
0.0312	COLING'04	Determining The Sentiment Of Opinions
0.0307	HLT'05	Extracting Product Features And Opinion From Reviews
0.0287	EMNLP'03	Towards Answering Opinion Questions: Separating Facts From Opinions And Identifying The Polarity Of Opinion Sentences
0.0279	EMNLP'03	Learning Extraction Patterns For Subjective Expressions
0.0169	ACL'05	Seeing Stars: Exploiting Class Relationships For Sentiment Categorization With Respect To Rating Scales

Table 3: Topic milestone papers (top-10 papers) for *Sentiment Analysis* in our collective model.

$\phi(c_i; z_k)$	Venue	Paper Title
0.1317	EMNLP'02	Thumbs Up? Sentiment Classification Using Machine Learning Techniques
0.0747	ACL'04	A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based On Minimum Cuts
0.0689	ACL'02	Thumbs Up Or Thumbs Down? Semantic Orientation Applied To Unsupervised Classification Of Reviews
0.0482	HLT'05	Recognizing Contextual Polarity In Phrase-Level Sentiment Analysis
0.0351	ACL'05	Seeing Stars: Exploiting Class Relationships For Sentiment Categorization With Respect To Rating Scales
0.0304	ACL'07	Biographies, Bollywood, Boom-boxes and Blenders: Domain Adaptation for Sentiment Classification
0.0215	ACL'07	Structured Models for Fine-to-Coarse Sentiment Analysis
0.0210	EMNLP'08	Learning with Compositional Semantics as Structural Inference for Subsentential Sentiment Analysis
0.0195	EMNLP'08	Multilingual Subjectivity Analysis Using Machine Translation
0.0177	ACL-IJCNLP'09	Co-Training for Cross-Lingual Sentiment Classification

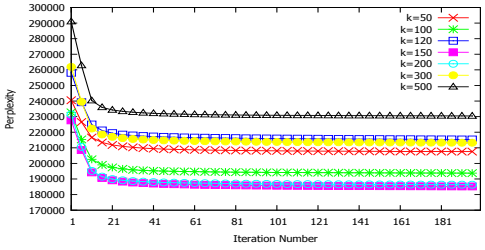

Figure 2: The perplexity over different k for our model

$$p(z_k|c_i, v_m) = \frac{\phi(c_i; z_k)\psi(z_k; v_m)}{\sum_k \phi(c_i; z_k)\psi(z_k; v_m)}$$

In the first E-step, the posterior probabilities are randomly initialized. The M-step that follows each E-step, re-estimates the θ which maximizes Q as follows:

$$\delta(z_k; d_j) = \frac{\sum_i N_{ij} p(z_k|c_i, d_j)}{\sum_i N_{ij}} \quad \zeta(z_k; a_n) = \frac{\sum_i U_{in} p(z_k|c_i, a_n)}{\sum_i N_{in}}$$

$$\psi(z_k; v_m) = \frac{\sum_i U_{im} p(z_k|c_i, v_m)}{\sum_i N_{im}}$$

$$\phi(c_i; z_k) \propto \alpha \frac{\sum_j N_{ij} p(z_k|c_i, d_j)}{\sum_i \sum_j N_{ij} p(z_k|c_i, d_j)} + \beta \frac{\sum_n U_{in} p(z_k|c_i, a_n)}{\sum_i \sum_n U_{in} p(z_k|c_i, a_n)} + \gamma \frac{\sum_m E_{im} p(z_k|c_i, v_m)}{\sum_i \sum_m E_{im} p(z_k|c_i, v_m)}$$

4. EXPERIMENTS

4.1 Dataset

The ACL Anthology Network (ANN) [7] was used in our experiments. There are 19408 papers by 15824 authors, published in 342 venues and having received 85367 citations. This dataset is also used in previous work [8]; thus, we can use it to perform some comparisons with [8].

Before testing our model, we have to determine the number of topics k. Perplexity is commonly used to evaluate performance of topic modeling. Thus, we select the value of k which minimizes perplexity. Figure 2 shows the perplexity scores during model estimation for different values of k. From this graph, we can see that a value of k around 150 is appropriate for this dataset, since it gives the lowest perplexity score among all tested values. Therefore, we perform our experiments using $k = 150$. In addition, we consider the three-typed documents equally and set the weights $\alpha = \beta = \gamma = 1$.

4.2 Experimental Results

4.2.1 Results of Topic Milestone Paper Discovery

Each topic is presented as the mixture of citations in our model. Within each topic z_k, $\phi(c_i; z_k)$ indicates the importance of citation (i.e., the cited paper) c_i. Those citations can be ranked based on $\phi(c_i; z_k)$ and citations ranking at the top for each topic z_k are considered as topic milestone papers. In order to compare with previous work [8], we use the top-10 papers for the topic *Sentiment Analysis* as an example. Table 2 presents topic milestone papers for *Sentiment Analysis* in [8] while Table 3 shows our results. There are 5 overlapping papers and 5 different papers between the two models. The top-1 paper is the same, but the order of the overlapping papers is slightly different. The top-2 paper in our model is ranked highly, compared with that in [8] (top-4). We have more recently published papers. The reason behind these differences is that the previous model does not consider factors such as published venues and authorship that have influence on paper importance.

4.2.2 Results of Venue Milestone Paper Discovery

In previous work, only topic milestone paper discovery has been studied. Our model is more general in the sense that it can identify milestone papers also for a given venue or author. In the next experiment, the probability of a citation given a venue is computed by Equation 3 and papers are ranked based on $p(c|v)$. Here we only take the venue

Table 4: Milestone papers (top-10) for *ACL* in our collective model.

$p(c\|v)$	Paper Title	Venue
0.0084	Building A Large Annotated Corpus Of English: The Penn Treebank	JCL'93
0.0074	The Mathematics Of Statistical Machine Translation: Parameter Estimation	JCL'93
0.0059	Statistical Phrase-Based Translation	NAACL'03
0.0057	Bleu: A Method For Automatic Evaluation Of Machine Translation	ACL'02
0.0056	A Hierarchical Phrase-Based Model For Statistical Machine Translation	ACL'05
0.0056	A Systematic Comparison Of Various Statistical Alignment Models	JCL'03
0.0050	Minimum Error Rate Training In Statistical Machine Translation	ACL'03
0.0050	A Maximum-Entropy-Inspired Parser	NAACL'00
0.0044	Stochastic Inversion Transduction Grammars And Bilingual Parsing Of Parallel Corpora	JCL'97
0.0044	Accurate Unlexicalized Parsing	ACL'03

Table 5: Author milestone papers (top-10) for the author *Bo Pang*.

$p(c\|a)$	Paper Title	Venue
0.0461*	Thumbs Up? Sentiment Classification Using Machine Learning Techniques	EMNLP'02
0.0375	Thumbs Up Or Thumbs Down? Semantic Orientation Applied To Unsupervised Classification Of Reviews	ACL'02
0.0253*	A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based On Minimum Cuts	ACL'04
0.0241	Predicting The Semantic Orientation Of Adjectives	ACL'97
0.0172	Extracting Product Features And Opinions From Reviews	HLT'05
0.0152	Towards Answering Opinion Questions: Separating Facts From Opinions And Identifying The Polarity Of Opinion Sentences	EMNLP'03
0.0147	Learning Extraction Patterns For Subjective Expressions	EMNLP'03
0.0093*	Seeing Stars: Exploiting Class Relationships For Sentiment Categorization With Respect To Rating Scales	ACL'05
0.0086	Biographies, Bollywood, Boom-boxes and Blenders: Domain Adaptation for Sentiment Classification	ACL'07
0.0065	Learning Subjective Language	JCL'04

ACL as example and Table 4 reports important papers in ACL. These papers are mainly from three venues, e.g ACL, NAACL and JCL (Journal of Computational Linguistics).

4.2.3 Results of Author Milestone Paper Discovery

Similarly, author milestone papers can be ranked, based on the probability of a citation given an author $p(c|a)$ (see Equation 2). We used the author *Bo Pang* (the first author of the top-1 paper in Table 3) as an example. Table 5 shows the citations that *Bo Pang* has the highest probability to cite (* indicates self-citation). The result has high overlap with Table 2 (8 papers). This might indicate that author citation patterns are regular and the author factor has similar influence as the citation relationships.

4.2.4 Topic Involvements

Finally, our model can also indicate popular topics for an author or a venue. $\psi(z; v)$ indicates how well a topic z represents a venue v and $\zeta(z; a)$ shows how well a topic z can represent an author a. Therefore, we can find the most popular topics for a specific venue or an author. The top-3 topics for ACL are respectively *Name entity extraction*, *Statistical parsing* and *Statistical machine translation*. The top-3 topics for *Bo Pang* are *Sentiment analysis*, *Opinion extraction* and *Paraphrases generation*.

5. CONCLUSIONS AND FUTURE WORK

In this paper, we proposed a collective topic model for multiple-typed objects in the academic network, in order to address the issue of milestone paper discovery. The model is based on PLSA, and authorship, published venues and citation relations have been included in it. Our method can not only discover topic milestone papers discussed in previous work, but also explore venue milestone papers and author milestone papers. In addition, it can find representative topics for an author or a venue. Experiments on a real dataset ANN show that our model can better evaluate the impact of papers and its result is not biased against new publications. Directions for future work include the investigation of more complicated models with biased mechanisms and the integration of this model into existing academic literature search/recommendation systems.

6. ACKNOWLEDGMENTS

This work was supported by grant HKU 715711E from Hong Kong RGC. The authors would like to thank the anonymous reviewers for their valuable comments.

7. REFERENCES

[1] S. Bethard and D. Jurafsky. Who should i cite: learning literature search models from citation behavior. In *CIKM '10*, pages 609–618. ACM, 2010.

[2] C. M. Bishop. *Pattern Recognition and Machine learning*. Springer, 2006.

[3] H. Deng, B. Zhao, and J. Han. Collective topic modeling for heterogeneous networks. In *SIGIR '11*, pages 1109–1110, New York, NY, USA, 2011. ACM.

[4] T. Hofmann. Probabilistic latent semantic indexing. In *SIGIR '99*, pages 50–57, New York, NY, USA, 1999. ACM.

[5] W. Huang, S. Kataria, C. Caragea, P. Mitra, C. L. Giles, and C. L. Giles. Recommending citations: translating papers into references. In *CIKM '12*, pages 1910–1914. ACM, 2012.

[6] Q. Mei and C. Zhai. Discovering evolutionary theme patterns from text: An exploration of temporal text mining. In *KDD '05*, pages 198–207, New York, NY, USA, 2005. ACM.

[7] D. R. Radev, P. Muthukrishnan, V. Qazvinian, and A. Abu-Jbara. The acl anthology network corpus. *Language Resources and Evaluation Journal*, pages 1–26, 2013.

[8] X. W. Wang, C. Zhai, and D. Roth. Understanding evolution of research themes: a probabilistic generative model for citations. In *KDD'13*, pages 1115–1123. ACM, 2013.