Are School-of-thought Words Characterizable?

Xiaorui Jiang[¶1] **Xiaoping Sun**[¶2] **Hai Zhuge**[¶†‡3*]

[¶] Key Lab of Intelligent Information Processing, Institute of Computing Technology, CAS, Beijing, China
[†] Nanjing University of Posts and Telecommunications, Nanjing, China
[‡] Aston University, Birmingham, UK

[1] xiaoruijiang@gmail.com [2] sunxp@kg.ict.ac.cn
[3] zhuge@ict.ac.cn

Abstract

School of thought analysis is an important yet not-well-elaborated scientific knowledge discovery task. This paper makes the first attempt at this problem. We focus on one aspect of the problem: do characteristic school-of-thought words exist and whether they are characterizable? To answer these questions, we propose a probabilistic generative School-Of-Thought (SOT) model to simulate the scientific authoring process based on several assumptions. SOT defines a school of thought as a distribution of topics and assumes that authors determine the school of thought for each sentence before choosing words to deliver scientific ideas. SOT distinguishes between two types of school-of-thought words for either the general background of a school of thought or the original ideas each paper contributes to its school of thought. Narrative and quantitative experiments show positive and promising results to the questions raised above.

1 Introduction

With more powerful computational analysis tools, researchers are now devoting efforts to establish a "science of better science" by analyzing the ecosystem of scientific discovery (Goth, 2012). Amongst this ambition, school of thought analysis has been identified an important fine-grained scientific knowledge discovery task. As mentioned by Teufel (2010), it is important for an experienced scientist to know which papers belong to which *school of thought* (or technical route) through years of knowledge accumulation. Schools of thought typically emerge with the evolution of a research domain or scientific topic.

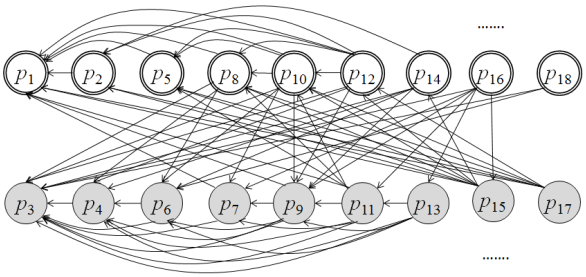

Figure 1. The citation graph of the reachability indexing domain (c.f. the RE data set in Table 1).

Take reachability indexing for example, which we will repeatedly turn to later, there are two schools of thought, the *cover-based* (since about 1990) and *hop-based* (since the beginning of the 2000s) methods. Most of the following works belong to either school of thought and thus two streams of innovative ideas emerge. Figure 1 illustrates this situation. Two chains of subsequentially published papers represent two schools of thought of the reachability indexing domain. The top chain of white double-line circles and the bottom chain of shaded circles represent the cover-based and hop-based streams respectively.

However it is not easy to gain this knowledge about school of thought. Current citation indexing services are not very helpful for this kind of knowledge discovery tasks. As explained in Figure 1, papers of different schools of thought cite each other heavily and form a rather dense citation graph. An extreme example is p_{14}, which cites more hop-based papers than its own school of thought.

If the current citation indexing service can be equipped with school of thought knowledge, it will help scientists, especially novice researchers, a lot in grasping the core ideas of a scientific domain quickly and making their own way of innovation (Upham et al., 2010). School of thought analysis is also useful for knowledge

* Corresponding author.

Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, pages 822–828, Sofia, Bulgaria, August 4-9 2013.

flow discovery (Zhuge, 2006; Zhuge, 2012), knowledge mapping (Chen, 2004; Herrera et al., 2010) and scientific paradigm summarization (Joang and Kan, 2010; Qazvinian et al., 2013) etc.

This paper makes the first attempts to unsupervised school of thought analysis. Three main aspects of school of thought analysis can be identified: determining the number of schools of thought, characterizing school-of-thought words and categorizing papers into one or several school(s) of thought (if applicable). This paper focuses on the second subproblem and leaves the other two as future work. Particularly, we purpose to investigate whether characteristic school-of-thought words exist and whether they can be automatically characterized. To answer these questions, we propose the probabilistic generative School-Of-Thought model (SOT for short) based on the following assumptions on the scientific authoring process.

Assumption A1. The co-occurrence patterns are useful for revealing which words and sentences are school-of-thought words and which schools of thought they describe. Take reachability indexing for example, hop-based papers try to get the "**optimum labeling**" by finding the "**densest intermediate hops**" to encode reachability information captured by an intermediate data structure called "**transitive closure contour**". To accomplish this, they solve the "**densest subgraph problem**" on specifically created "**bipartite**" graphs centered at "**hops**" by transforming the problem into an equivalent "**minimum set cover**" framework. Thus, these bold-faced words often occur as hop-based school-of-thought words. In cover-based methods, however, one or several "**spanning tree(s)**" are extracted and "**(multiple) intervals**" are assigned to each node as reachability labels by "**pre-(order)**" and "**post-order traversals**". Meanwhile, graph theory terminologies like "**root**", "**child**" and "**ancestor**" etc. also frequently occur as cover-based school-of-thought words.

Assumption A2. Before writing a sentence to deliver their ideas, the authors need to determine which school of thought this sentence is to portray. This is called the *one-sot-per-sentence* assumption, where "sot" abbreviates "school of thought". The one-sot-per-sentence assumption does not mean that authors intentionally write this way, but only simulates the outcome of the scientific paper organization. Investigations into scientific writing reveal that sentences of different schools of thought can occur anywhere and are often interleaved. This is because authors of a scientific paper not only contribute to the school of thought they follow but also discuss different

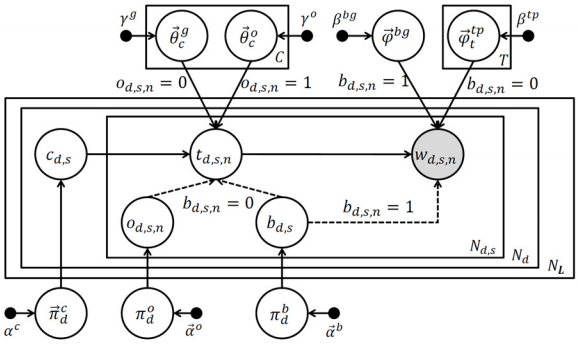

Figure 2. The SOT Model

schools of thought. For example, in the **Method** part, the authors may turn to discuss another paper (possibly of a different school of thought) for comparison. This phenomenon also occurs frequently in the **Results** or **Discussions** section. Besides, citation sentences often acknowledge related works of different schools of thought.

Assumption A3. All the papers of a domain talk about the general domain backgrounds. For example, reachability indexing aims to build "**compact indices**" for facilitating "**reachability queries**" between "**source**" and "**target nodes**". Other background words include "**(complete) transitive closure**", "**index size**" and "**reach**" etc., as well as classical graph theory terminologies like "**predecessors**" and "**successors**" etc.

Assumption A4. Besides contributing *original* ideas, papers of the same school of thought typically need to follow some *general* strategies that make them fall into the same school of thought. For example, all the hop-based methods follow the general ideas of designing approximate algorithms for choosing good hops, while the original ideas of each paper lead to different labeling algorithms. Scientific readers pay attention to the original ideas of each paper as well as the general ideas of each school of thought. This assumes that a word can be either a *generality* or *originality* word to deliver general and original ideas of a school of thought respectively.

2 The School-of-Thought Model

Figure 2 shows the proposed SOT model. SOT reflects all the assumptions made in Sect. 1. The plate notation follows Bishop (2006) where a shaded circle means an observed variable, in this context word occurrence in text, a white circle denotes either a latent variable or a model parameter, and a small solid dot represents a hyper-parameter of the corresponding model parameter. The generative scientific authoring process illustrated in Figure 2 is elaborated as follows.

Step 1. School of thought assignment (A2).

Data Sets	N_L	W	S	N_d (avg)	C	Schools of Thought (Number of Papers under this School of Thought)
RE	18	54035	5300	294	2	Hop-Based (9), Cover-Based (9)
NP	24	36227	3329	138	3	Mention-Pair Models (14), Entity-Mention Models (5), Ranking Models (5)
PP	20	21941	2182	109	4	Using Single Monolingual Corpus (3), Using Monolingual Parallel Corpora (6), Using Monolingual Comparable Corpora (5), Using Bilingual Parallel Corpora (5)
TE	34	55671	5335	156	2	Finite-State Transducer models (17), Synchronous Context-Free Grammar models (17)
WA	18	19219	1807	100	3	Asymmetric Models (5), Symmetric Alignment Models (9), Supervised Learning for Alignment (4)
DP	56	68384	6021	107	3	Transition-Based (20), Graph-Based (17), Grammar-Based (19)
LR	44	77024	7395	168	3	Point-wise Approach (11), Pair-wise Approach (17), List-wise Approach (16)

Notes: RE – REachability indexing; NP – Noun Phrase co-reference resolution; PP – ParaPhrase; TE – Translational Equivalence; WA – Word Alignment; DP – Dependency Parsing; LR – Learning to Rank; W – number of words; S – number of sentences; C – gold-standard number of schools of thought; N_d – number of sentences in document d.

Table 1. Data Sets

To simulate the one-sot-per-sentence assumption, we introduce a latent school-of-thought assignment variable $c_{d,s}$ ($1 \le c_{d,s} \le C$, where C is the number of schools of thought) for each sentence s in paper d, dependent on which are topic assignment and word occurrence variables. As different papers and their authors have different foci, flavors and writing styles, it is appropriate to assume that each paper d has its own Dirichlet distribution of schools of thought $\vec{\pi}_d^c \sim Dir(\vec{\alpha}^c)$ (refer to Heinrich (2008) for Dirichlet analysis of texts). $c_{d,s}$ is thus multinomially sampled from $\vec{\pi}_d^c$, that is, $c_{d,s} \sim Mutl(\vec{\pi}_d^c)$.

Step 2. Background word emission (A3).

Before choosing a word $w_{d,s,n}$ to deliver scientific ideas, the authors first need to determine whether this word describes domain backgrounds or depicts a specific school-of-thought. This information is indicated by the latent background word indicator variable $b_{d,s,n} \sim Bern(\pi_d^b)$, where $\pi_d^b \sim Beta(\alpha_0^b, \alpha_1^b)$ is the probability of Bernoulli test. $b_{d,s,n} = 1$ means $w_{d,s,n}$ is a background word that is multinomially sampled from the Dirichlet background word distribution $\vec{\varphi}^{bg} \sim Dir(\vec{\beta}^{bg})$, i.e. $w_{d,s,n} \sim Mutl(\vec{\varphi}^{bg})$.

Step 3. Originality indicator assignment (A4).

If $b_{d,s,n} = 0$, $w_{d,s,n}$ is a school-of-thought word. Then the authors need to determine whether $w_{d,s,n}$ talks about the general ideas of a certain school of thought (i.e. a *generality* word when $o_{d,s,n} = 0$) or delivers original contributions to the specific school of thought (i.e. an *originality* word when $o_{d,s,n} = 1$). The latent originality indicator variable $o_{d,s,n}$ is assigned in a similar way to $b_{d,s,n}$.

Step 4. Topical word emission.

SOT regards schools of thought and topics as two different levels of semantic information. A school of thought is modeled as a distribution of topics discussed by the papers of a research domain. Each topic in turn is defined as a distribution of the topical words. Reflected in Figure 1, $\vec{\theta}_c^g$ and $\vec{\theta}_c^o$ are Dirichlet distributions of generality and originality topics respectively, with γ^g and γ^o being the Dirichlet priors. According to the assignment of the originality indicator, the topic $t_{d,s,n}$ of the current token is multinomially selected from either $\vec{\theta}_c^g$ ($o_{d,s,n} = 0$) or $\vec{\theta}_c^o$ ($o_{d,s,n} = 1$). After that, a word $w_{d,s,n}$ is multinomially emitted from the topical word distribution $\vec{\varphi}_{t_{d,s,n}}^{tp}$, where $\vec{\varphi}_t^{tp} \sim Dir(\beta^{tp})$ for each $1 \le t \le T$.

Gibbs sampling is used for SOT model inference. Considering the logic of presentation, it is detailed in Appendix B.

3 Experiments

3.1 Datasets

Lacking standard test benchmarks, we compiled 7 data sets according to well-known recent surveys (see Appendix A). Each data set consists of several dozens of papers of the same domain. When constructing these data sets, the only place of human intervention is the de-duplication step, which means typically only one of a number of highly duplicated references is kept in the data set. Different from previous studies reviewed in Sect. 4, full texts but not abstracts are used. We extracted texts from the collected papers and removed tables, figures and sentences full of math equations or unrecognizable symbols. The statistics of the resulting data sets are listed in Table 1. The gold-standard number and the classification of schools of thoughts reflect not only the viewpoints of the survey authors but also the consensus of the corresponding research communities.

3.2 Qualitative Results

This section looks at the capabilities of SOT in learning background and school-of-thought words using the RE data set as an example. Given the estimated model parameters, the distributions of the school-of-thought words of SOT can be calculated as weighted sums of topical word emission probabilities ($\varphi_{t,w}^{tp}$ for each word w) over all the topics ($\sum_t$) and papers ($\sum_d$), as in Eq. (1).

BACKGROUND WORDS			SCHOOL-OF-THOUGHT WORDS					
			SOT-1 (COVER-BASED)			SOT-2 (HOP-BASED)		
node	**arc**	**figure**	node	reachable	find	**2-hop**	**problem**	**hop**
closure	**size**	deleted	graph	reach	reachability	vertex	tree	**subgraph**
chain	lists	incremental	nodes	size	***cover***	vertices	edges	proposed
graph	procedure	**predecessor**	closure	**chains**	acyclic	***cover***	construction	large
nodes	arcs	directed	**tree**	graphs	database	graph	**approach**	**path-hop**
compressed	update	**edge**	edges	storage	**traversal**	algorithm	indexing	**lin**
list	off-chain	systems	chain	instance	components	size	**contour**	spanning
transitive	**acyclic**	**connected**	transitive	**intervals**	directed	chain	processing	smaller
successor	**reduction**	techniques	**non-tree**	**spanning**	lists	**labeling**	chain	**optimal**
compression	relation	single	number	segment	reduction	closure	pairs	**densest**
storage	**source**	**cycles**	compressed	**order**	g.	reachability	**compression**	**decomposition**
chains	**reach**	updates	**path**	connected	addition	transitive	reachable	dag
required	effort	depth	edge	component	technique	graphs	property	paths
index	**obtained**	**materialize**	index	case	degree	time	figure	data
number	**component**	concatenation	list	**postorder**	gs	number	path-tree	**ratio**
database	**path**	presented	***set***	strongly	successors	**3-hop**	**bipartite**	nodes
case	**assignment**	added	**interval**	original	**structure**	index	**scheme**	edge
technique	**predecessors**	original	successor	**ris**	single	**labels**	**density**	**finding**
degree	addition	**components**	figure	required	paths	query	queries	**rank**
successors	**indices**	**strongly**	compression	source	arc	***set***	reach	note
destination (65), **determine** (76), **pair** (77), **resulting** (84), **merging** (86), **reached** (87), **store** (96)			**root** (67), **pre-** (85), **topological** (96), **sub-tree** (102), **ancestor** (105), **child** (106), **multiple** (113), **preorder** (117)			**lout** (66), **segment** (68), **minimum** (69), **in-termediate** (77), **greedy** (87), **faster** (88), **heuristics** (92), **approximate** (120)		

Table 2. The distributions of top-120 background and school-of-thought words.

$$p(w \mid c, o = 0/1) = \sum_d \left(\frac{N_{d,v}(d,w)}{N_v(w)} \pi^o_{d,0/1} \sum_t \theta^{g/o}_{c,t} \varphi^{tp}_{t,w} \right) \quad (1)$$

The first row of Table 2 lists the top-60 background and school-of-thought words learned by SOT for the RE data set sorted in descending order of their probabilities column by column. The words at the bottom are some of the remaining characteristic words together with their positions on the top-120 list. In the experiments, T is set to 20. As the data sets are relative small, it is not appropriate to set T too large, otherwise most of the topics are meaningless or duplicate. Either case will impose additive negative influences on the usefulness of the model, for example when applied to schools of thought clustering in the next section. C is set to the gold-standard number of schools of thought as in this study we are mainly interested in whether school-of-thought words are characterizable. The problems of identifying the existence and number of schools of thought are left to future work. Other parameter settings follow Griffiths and Steyvers (2010). The learned word distributions are shown very meaningful at the first glance. They are further explained as follows.

For domain backgrounds, reachability indexing is a classical problem of the graph database "**domain**" which talks about the reachability between the "**source**" and "**destination nodes**" on a "**graph**". Reachability "**index**" or "**indices**" aim at a "**reduction**" of the "**transitive closure**" so as to make the "**required storage**" smaller. All current works preprocess the input graphs by "**merging strongly connected components**" into representative nodes to remove "**cycles**".

We then give a deep investigation into the hop-based school-of-thought words (SoT-2). Cover-based ones conform well to the assumptions in Sect. 1 too. "**2-hop**", "**3-hop**" and "**path-hop**" are three representative hop-based reachability "**labeling schemes**" (a phrase preferred by hop-based papers). Hop-based methods aim at "**finding**" the "**optimum labeling**" with "**minimum cost**" and achieving a higher "**compression ratio**" than cover-based methods. To accomplish this, hop-based methods define a "**densest subgraph problem**" on a "**bipartite**" graph, transform it to an equivalent "**set cover**" problem, and then apply "**greedy**" algorithms based on several "**heuristics**" to find "**approximate**" solutions. The "**intermediate hops**" with the highest "**density**" are found as labels and assigned to "L_{out}" and "L_{in}" of certain "**contour**" vertices. "**contour**" is used by hop-based methods as a concise representation of the remaining to-be-encoded reachability information.

The underlined bold italic words such as "***set***" and "***cover***" are misleading (yet not necessarily erroneous) words as both schools of thought use them heavily, but in quite different contexts, for example, a "***set***" of labels versus "***set cover***", and "***cover(s)***" partial reachability information versus tree "***cover***". To improve, one of our future works shall integrate multi-word expressions or n-grams (Wallach, 2006) and syntactic analysis (Griffiths et al., 2004) into the current model.

3.3 Quantitative Results

To see the usefulness of school-of-thought words, we use the SOT model as a way to feature space reduction for a more precise text representation in the school-of-thought clustering task. A subset of school-of-thought words whose accumulated probability exceeds a given threshold *fsThr* are used as the reduced feature vector. Text is represented in the vector space model weighted using *tf*·*idf*. *K*-means is used for clustering. To obtain a stable and reliable result, we choose 300 random seeds as initial cluster centroids, run *K*-means 300 times and, following the heuristic suggestion by Manning et al. (2009), output the best clustering by the minimum residual squared sum principle. Two baselines are the "RAW" method without dimension reduction and LDA-based (Blei et al., 2003) feature selection. Table 3 reports the *F*-measure values of different competitors. In the parentheses are the corresponding threshold values under which the reported clustering result is obtained. The larger the threshold value is, the less effective the method in dimension reduction.

Compared to the baselines, SOT has consistently the best clustering qualities. When $fsThr \leq 0.70$, the feature space is reduced from several thousand words to only a few hundreds. LDA is typically better than RAW (except on LR) but less efficient in dimension reduction, e.g. on WA and DP. In the latter two cases, $fsThr = 0.80$ typically means LDA is much less efficient in feature reduction than SOT on these two data sets.

DATA SETS	F-MEASURE ($\beta = 2.0$)		
	RAW	LDA (*fsThr*)	SOT (*fsThr*)
RE	.7464	.7464 (.50)	**.7482** (.60)
NP	.4528	.6150 (.75)	**.6911** (.75)
PP	.3256	.4179 (.60)	**.6025** (.75)
TE	.2580	.5148(.60)	**.9405** (.40)
WA	.3125	.4569 (.80)	**.5519** (.60)
DP	.4787	.6762 (.80)	**.7155** (.50)
LR	.5413	*.5276* (*.95*)	**.6583** (.75)

Table 3. School-of-thought clustering results

4 Related Work

An early work in semantic analysis of scientific articles is Griffiths and Steyvers (2004) which focused on efficient browsing of large literature collections based on scientific topics. Other related researches include topic-based reviewer assignment (Mimno and McCallum, 2007), citation influence estimation (Dietz et al., 2007), research topic evolution (Hall et al., 2008) and expert finding (Tu et al., 2010) etc.

Another line of research is the joint modeling of topics and other types of semantic units such as perspectives (Lin et al., 2006), sentiment (Mei et al., 2007) and opinions (Zhao et al., 2010) etc. These works also took a multi-dimensional view of document semantics. The TAM model (Paul and Girju, 2010) might be the most relevant to SOT. TAM simultaneously models aspects and topics with different assumptions from SOT and it models purely on word level.

Studies that introduce an explicit background distribution include Chemudugunta et al. (2006), Haghighi and Vanderwende (2009), and Li et al. (2010) etc. Different from these works, SOT assumes that not only some "meaningless" general-purpose words but also more meaningful words about the specific domain backgrounds can be learned. What's more these works all model on a word level.

However, it is very useful to regard sentence as the basic processing unit, for example in the text scanning approach simulating human reading process by Xu and Zhuge (2013). Indeed, sentence-level school of thought assignment is crucial to SOT as it allows SOT to model the scientific authoring process. There are also other works that model text semantics on different levels other than words or tokens, such as Wallach (2006) on *n*-grams and Titov and McDonald (2008) on words within multinomially sampled sliding windows. The latter also distinguishes between different levels of topics, say global versus local topics, while in SOT such discrimination is generality versus originality topics.

5 Conclusion

This paper proposes a probabilistic generative model SOT for characterizing school-of-thought words. In SOT, a school of thought is modeled as a distribution of topics, with the latter defined as a distribution of topical words. School of thought assignment to each sentence is vital as it allows SOT to simulate the scientific authoring process in which each sentence conveys a piece of idea contributed to a certain school of thought as well as the domain backgrounds. Narrative and quantitative analysis show that high-quality school-of-thought words can be captured by the proposed model.

Acknowledgements

This work is partially supported by National Science Foundation of China (No. 61075074 and No. 61070183) and funding from Nanjing University of Posts and Telecommunications. Special thanks go to Prof. Jianmin Yao at Soochow University and Suzhou Scientific Service Center of China for his advices and suggestions that help this paper finally come true.

References

Chemudugunta, C., Smyth P., and Steyvers, M. 2006. Modeling general ad specific aspects of documents with a probabilistic topic model. In *Proc. NIPS'06*.

Bishop, C. M. 2006. *Patter Recognition and Machine learning*. Ch. 8 Graphical Models. Springer.

Blei, D. M., Ng, A. Y., and Jordan, M. I. 2003. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3: 993–1022.

Chen, C. 2004. Searching for intellectual turning points: Prograssive knowledge domain visualization. *Proc. Natl. Acad. Sci*., 101(suppl. 1): 5303–5310.

Dietz, L., Bickel, S., and Scheffer, T. 2007. Unsupervised prediction of citation influence. In *Proc. ICML'07*, 233–240.

Goth, G. 2012. The science of better science. *Commun. ACM,* 55(2): 13–15.

Griffiths, T., and Steyvers, M. 2004. Finding scientific topics. *Proc. Natl. Acad. Sci.*, 101 (suppl 1): 5228–5235.

Griffiths, T., Steyvers, M., Blei, D. M., and Tenenbaum, J. B. 2004. Integrating topics and syntax. In *Proc. NIPS'04*.

Haghighi, A., and Vanderwende, L. 2009. Exploring content models for multi-document summarization. In *Proc. HLT-NAACL'09*, 362–370.

Hall, D., Jurafsky, D., and Manning, C. D. 2008. Studying the history of ideas using topic models. In *Proc. EMNLP'08*, 363–371.

Heinrich, G. 2008. Parameter estimation for text analysis. Available at www.arbylon.net/publications/text-est.pdf.

Herrera, M., Roberts, D. C., and Gulbahce, N. 2010. Mapping the evolution of scientific fields. *PLoS ONE*, 5(5): e10355.

Joang, C. D. V., and Kan, M.-Y. (2010). Towards automatic related work summarization. In *Proc. COLING 2010*.

Li, P., Jiang, J., and Wang, Y. 2010. Generating templates of entity summaries with an entity-aspect model and pattern mining. In *Proc. ACL'10*, 640–649.

Lin, W., Wilson, T., Wiebe, J., and Hauptmann, A. 2006. Which side are you on? Identifying perspectives at the document and sentence levels. In *Proc. CoNLL'06*, 109–116.

Manning, C. D., Raghavan, P., and Schütze, H. 2009. *Introduction to Information Retrieval*. Ch. 16. Flat Clustering. Cambridge University Press.

Mei, Q., Ling, X., Wondra, M., Su, H., and Zhai, C. 2007. Topic sentiment mixture: modeling facets and opinions in weblogs. In *Proc. WWW'07*, 171–180.

Mimno, D., and McCallum, A. 2007. Expertise modeling for matching papers with reviewers. In *Proc. SIGKDD'07*, 500–509.

Paul, M., and Girju, R. 2010. A two-dimensional topic-aspect model for discovering multi-faceted topics. In *Proc. AAAI'10*, 545–550.

Qazvinian, V., Radev, D. R., Mohammad, S. M., Dorr, B., Zajic, D., Whidby, M., and Moon T. (2013). Generating extractive summaries of scientific paradigms. *J. Artif. Intell. Res.*, 46: 165–201.

Teufel, S. 2010. *The Structure of Scientific Articles.* CLSI Publications, Stanford, CA, USA.

Titov, I., and McDonald R. 2008. Modeling online reviews with multi-grain topic models. In *Proc. WWW'08*, 111–120.

Tu, Y., Johri, N., Roth, D., and Hockenmaier, J. 2010. Citation author topic model in expert search. In *Proc. COLING'10*, 1265–1273.

Upham, S. P., Rosenkopf, L., Ungar, L. H. 2010. Positioning knowledge: schools of thought and new knowledge creation. *Scientometrics*, 83 (2): 555–581.

Wallach, H. 2006. Topic modeling: beyond bag-of-words. In *Proc. ICML'06*, 977– 984.

Xu, B., and Zhuge, H. 2013. A text scanning mechanism simulating human reading process, In *Proc. IJCAI'13*.

Zhao, X., Jiang, J., Yan, H., and Li, X. 2010. Jointly modeling aspects and opinions with a MaxEnt-LDA hybrid. In *Proc. EMNLP'10*, 56– 65.

Zhuge, H. 2006. Discovery of knowledge flow in science. *Commun. ACM*, 49(5): 101-107.

Zhuge, H. 2012. *The Knowledge Grid: Toward Cyber-Physical Society* (2nd edition). World Scientific Publishing Company, Singapore.

Appendices

A Survey Papers for Building Data Sets

[RE] Yu, P. X., and Cheng, J. 2010. *Managing and Mining Graph Data*, Ch. 6, 181–215. Springer.

[NP] Ng, V. 2010. Supervised noun phrase coreference research: The first fifteen years. In *Proc. ACL'10*, 1396–1141.

[PP] Madnani, N., and Dorr, B. J. 2010. Generating phrasal and sentential paraphrases: A survey of data-driven methods. *Comput. Linguist.*, 36 (3): 341–387.

[TE/WA] Lopez, A. 2008. Statistical machine translation. *ACM Comput. Surv.*, 40(3), Article 8, 49 pages.

[DP] Kübler, S., McDonald, R., and Nivre, J. 2009. *Dependency parsing*, Ch. 3–5, 21–78. Morgan & Claypools Publishers.

[LR] Liu, T. Y. 2011. *Learning to rank for information retrieval*, Ch. 2–4, 33–88. Springer.

B Gibbs Sampling of the SOT Model

Using collapsed Gibbs sampling (Griffiths and Steyvers, 2004), the latent variable $\vec{c}$ is inferenced in Eq. (B1). In Eq. (B1), $N_{c,b,o,t}(c,0,o,t)$

$$p(c_{d,s}=c\mid \vec{c}^{\neg(d,s)},\bullet)\propto \prod_{t=1}^{T}\frac{\Gamma(N_{c,b,o,t}(c,0,0,t)+\gamma^{g})}{\Gamma(N_{c,b,o,t}^{\neg(d,s)}(c,0,0,t)+\gamma^{g})}\times\frac{\Gamma(N_{c,b,o,t}^{\neg(d,s)}(c,0,0,\Sigma)+T\cdot\gamma^{g})}{\Gamma(N_{c,b,o,t}(c,0,0,\Sigma)+T\cdot\gamma^{g})}$$
$$\times\prod_{t=1}^{T}\frac{\Gamma(N_{c,b,o,t}(c,0,1,t)+\gamma^{o})}{\Gamma(N_{c,b,o,t}^{\neg(d,s)}(c,0,1,t)+\gamma^{o})}\times\frac{\Gamma(N_{c,b,o,t}^{\neg(d,s)}(c,0,1,\Sigma)+\gamma^{o})}{\Gamma(N_{c,b,o,t}(c,0,1,\Sigma)+\gamma^{o})}\times\frac{N_{d,c}^{\neg(d,s)}(d,c)+\alpha^{c}}{N_{d,c}^{\neg(d,s)}(d,\Sigma)+C\cdot\alpha^{c}} \tag{B1}$$

$$p(b_{d,s,n}=1\mid w_{d,s,n}=v,\bullet)\propto\frac{N_{d,b}^{\neg(d,s,n)}(d,1)+\alpha_{1}^{b}}{N_{d,b}^{\neg(d,s,n)}(d,\Sigma)+\alpha_{0}^{b}+\alpha_{1}^{b}}\times\frac{N_{b,v}^{\neg(d,s,n)}(1,v)+\beta^{bg}}{N_{b,v}^{\neg(d,s,n)}(1,\Sigma)+V\cdot\beta^{bg}} \tag{B2}$$

$$p(b_{d,s,n}=0,o_{d,s,n}=0,t_{d,s,n}=t\mid c_{d,s}=c,\vec{b}^{\neg(d,s,n)},\vec{o}^{\neg(d,s,n)},\vec{t}^{\neg(d,s,n)},w_{d,s,n}=v,\bullet)$$
$$\propto\frac{N_{d,b}^{\neg(d,s,n)}(d,0)+\alpha_{0}^{b}}{N_{d,b}^{\neg(d,s,n)}(d,\Sigma)+\alpha_{0}^{b}+\alpha_{1}^{b}}\times\frac{N_{d,b,o}^{\neg(d,s,n)}(d,0,0)+\alpha_{0}^{o}}{N_{d,b,o}^{\neg(d,s,n)}(d,0,\Sigma)+\alpha_{0}^{o}+\alpha_{1}^{o}} \tag{B3}$$
$$\times\frac{N_{c,b,o,t}^{\neg(d,s,n)}(c,0,0,t)+\gamma^{g}}{N_{c,b,o,t}^{\neg(d,s,n)}(c,0,0,\Sigma)+T\cdot\gamma^{g}}\times\frac{N_{b,t,v}^{\neg(d,s,n)}(0,t,v)+\beta^{tp}}{N_{b,t,v}^{\neg(d,s,n)}(0,t,\Sigma)+V\cdot\beta^{tp}}$$

$$p(b_{d,s,n}=0,o_{d,s,n}=1,t_{d,s,n}=t\mid c_{d,s}=c,\vec{b}^{\neg(d,s,n)},\vec{o}^{\neg(d,s,n)},\vec{t}^{\neg(d,s,n)},w_{d,s,n}=v,\bullet)$$
$$\propto\frac{N_{d,b}^{\neg(d,s,n)}(d,0)+\alpha_{0}^{b}}{N_{d,b}^{\neg(d,s,n)}(d,\Sigma)+\alpha_{0}^{b}+\alpha_{1}^{b}}\times\frac{N_{d,b,o}^{\neg(d,s,n)}(d,0,1)+\alpha_{1}^{o}}{N_{d,b,o}^{\neg(d,s,n)}(d,0,\Sigma)+\alpha_{0}^{o}+\alpha_{1}^{o}} \tag{B4}$$
$$\times\frac{N_{c,b,o,t}^{\neg(d,s,n)}(c,0,1,t)+\gamma^{o}}{N_{c,b,o,t}^{\neg(d,s,n)}(c,0,1,\Sigma)+T\cdot\gamma^{o}}\times\frac{N_{b,t,v}^{\neg(d,s,n)}(0,t,v)+\beta^{tp}}{N_{b,t,v}^{\neg(d,s,n)}(0,t,\Sigma)+V\cdot\beta^{tp}}$$

Figure B1. The SOT model inference.

is the number of words of topic t describing the common ideas ($o = 0$) or original ideas ($o = 1$) of school of thought c. The superscript $\neg(d,s)$ means that words in sentence s of paper d are not counted. $N_{d,c}^{\neg(d,s)}(d,c)$) counts the number of sentences in paper d describing school of thought c with sentence s removed from consideration. In Eqs. (B1)–(B4), the symbol Σ means summation over the corresponding variable. For example,

$$N_{c,b,o,t}(c,0,o,\Sigma)=\sum\nolimits_{t=1,\cdots,T}N_{c,b,o,t}(c,0,o,t) \tag{B5}$$

Latent variables $\vec{b}$, $\vec{o}$ and $\vec{t}$ are jointly sampled in Eqs. (B2)–(B4). $N_{d,b}^{\neg(d,s,n)}(d,b)$ counts the number of background ($b = 0$) or school-of-thought ($b = 1$) words in document d without counting the n-th token in sentence s. $N_{b,v}^{\neg(d,s,n)}(1,v)$ is the number of times vocabulary item v occurs as background word in the literature collection without counting the n-th token in sentence s of paper d. $N_{d,b,o}^{\neg(d,s,n)}(d,0,o)$ is the number of words describing either common ideas ($o = 0$) or original ideas ($o = 1$) of some school of thought without considering the n-th token in sentence s of paper d. $N_{c,b,o,t}^{\neg(d,s,n)}(c,0,o,t)$ is the number of words of topic t in the literature collection describing either common ideas ($o = 0$) or original ideas ($o = 1$) of school of thought c without counting the n-th token in sentence s of paper d. $N_{b,t,v}^{\neg(d,s,n)}(0,t,v)$ is the number of school-of-thought words of topic t which is instantiated by vocabulary item v in the literature collection without counting the n-th token in sentence s of paper d.