Dynamic Topic Modeling of Systems Research Papers

Theano Stavrinos, Christopher Hodsdon

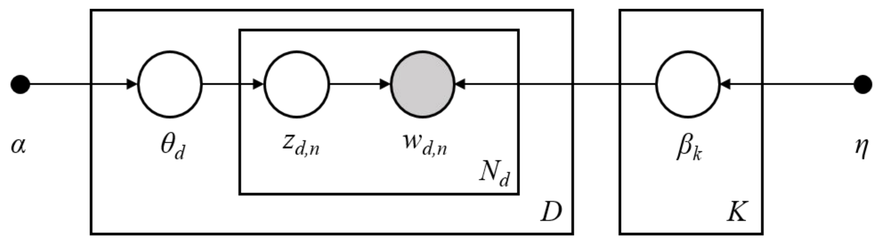
Princeton University

ABSTRACT

- The Symposium on Operating Systems Design and Implementation (OSDI) and the Symposium on Operating Systems Principles (SOSP) are the premiere conferences in the field of computer systems.
- The publications in these venues can give insight into evolution of computing systems over time, which is of historical interest.
- Topic analysis can also improve the searchability and categorization of conference papers by uncovering more detailed categories than manual labeling.
- We use a dynamic topic model [1] to analyze OSDI and SOSP topics over time.

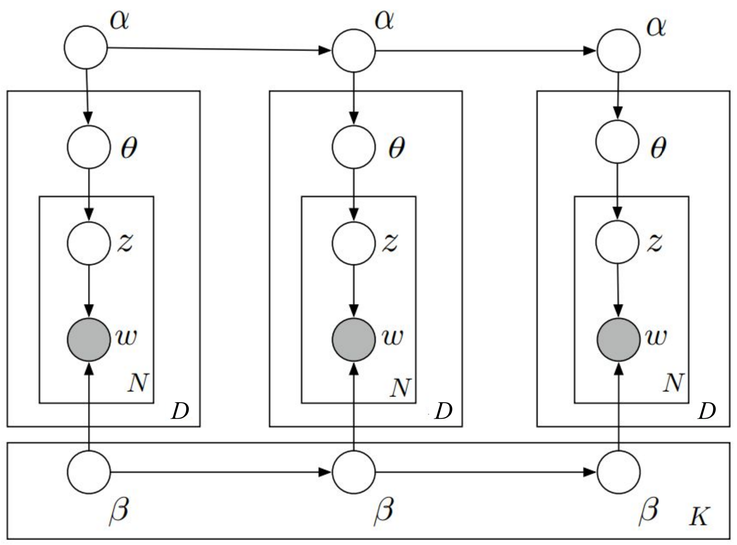
BACKGROUND AND APPROACH

Graphical representation of LDA



The graphical model for Latent Dirichlet Allocation [2]. α: Dirichlet distribution parameter to draw a document's topics. θ_d: per-document topic proportions. $Z_{d,n}$: the word's topic assignment. $W_{d,n}$: nth word for document d. η: hyperparameter for a Dirichlet distribution generating β_k, the set of topics and their word proportions. D: the number of documents. K: the number of topics. N_d: the number of words in document d.

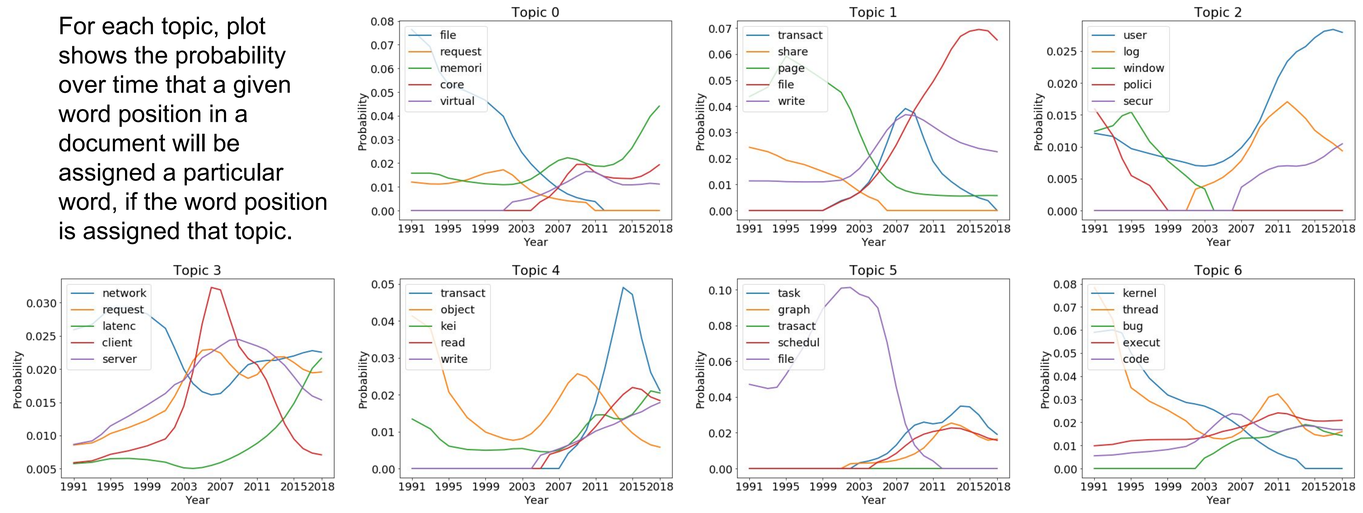
Graphical representation of the Dynamic Topic Model



The graphical model for Dynamic Topic Modeling. The parameter labels are the same as for LDA (above). Each vertical plate corresponds to a timepoint in the longitudinal analysis; prior word distributions per topic (β) and topic distributions per document (α) for each year in the model depend on the α and β from the previous timepoint.

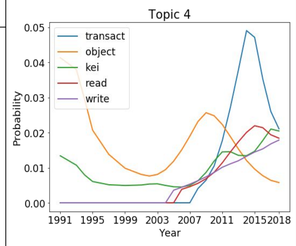
EVOLUTION IN WORD PROBABILITY PER TOPIC

For each topic, plot shows the probability over time that a given word position in a document will be assigned a particular word, if the word position is assigned that topic.

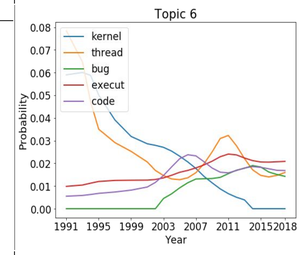


TOP 10 WORDS IN A TOPIC OVER TIME

1991	2005	2018
object	secure	transact
interface	code	key
key	packet	read
user	host	write
layer	user	store
type	object	throughput
code	network	hash
domain	address	value
service	label	client
call	control	server



1991	2005	2018
thread	kernel	execute
kernel	driver	program
user	code	code
processor	memory	thread
schedule	execute	failure
space	function	test
address	thread	bug
block	linux	analysis
execut	check	develop
interrupt	program	event



DOCUMENT SPOTLIGHT: MAPREDUCE TOPIC PROPORTIONS

- MapReduce [3] is a system for easily specifying and executing certain types of parallel computations on clusters
- Breaks computation into smaller 'tasks' transparently, masking complexity of distributed computing
- Networking component due to distributed nature of computations
- Hugely influential: 26K+ citations

Topic	Topic proportion
0 (memory/file systems)	0.046
1 (databases)	0.008
2 (security/usability)	**0.533**
3 (networking)	**0.270**
4 (distributed databases)	0.003
5 (analytics/scheduling)	**0.003**
6 (OS)	0.136

- Table: topic proportions (θ) for MapReduce paper
- Rows (topics) are labeled with interpreted real-world subjects
- High θ for security/usability and networking, as expected.
- Low θ for analytics, but topic 5 shows increased use of 'task' after 2004, suggesting influence on future work

FUTURE WORK

Papers from early (pre-'91) SOSP had poor PDF quality, which resulted in poor-quality text parsing from `pdftotext`. More work on compiling data could allow for analysis on papers as early as 1969. All analysis was done on unigrams; extending to bigrams or n-grams may give a deeper analysis: e.g., by distinguishing "file" and "file system".

REFERENCES

1. Blei, David M., and John D. Lafferty. "Dynamic topic models". In Proceedings of the 23rd international conference on Machine learning, pp. 113-120. ACM, 2006.
2. Blei, David M., Andrew Y. Ng, and Michael I. Jordan. "Latent dirichlet allocation." Journal of machine Learning research 3, no. Jan (2003): 993-1022.
3. Dean, Jeffrey, and Sanjay Ghemawat. "MapReduce: simplified data processing on large clusters." *Symposium on Operating Systems Design & Implementation* (2004).