The sensitivity of Latent Dirichlet Allocation for Information Retrieval

Laurence A. F. Park and Kotagiri Ramamohanarao
Department of Computer Science and Software Engineering
The University of Melbourne, Australia

The use of topic models adds to the precision of Document retrieval.

The use of topic models adds to the precision of Document retrieval.

Latent Dirichlet Allocation allows us to model documents with Dirichlet distributed topics.

The use of topic models adds to the precision of Document retrieval.

Latent Dirichlet Allocation allows us to model documents with Dirichlet distributed topics.

The document models depend on the Dirichlet parameter. Initial versions of LDA fit the parameter, while latter versions choose a value for the parameter.

The use of topic models adds to the precision of Document retrieval.

Latent Dirichlet Allocation allows us to model documents with Dirichlet distributed topics.

The document models depend on the Dirichlet parameter. Initial versions of LDA fit the parameter, while latter versions choose a value for the parameter.

How does this Dirichlet parameter affect the quality of Document retrieval?

Outline

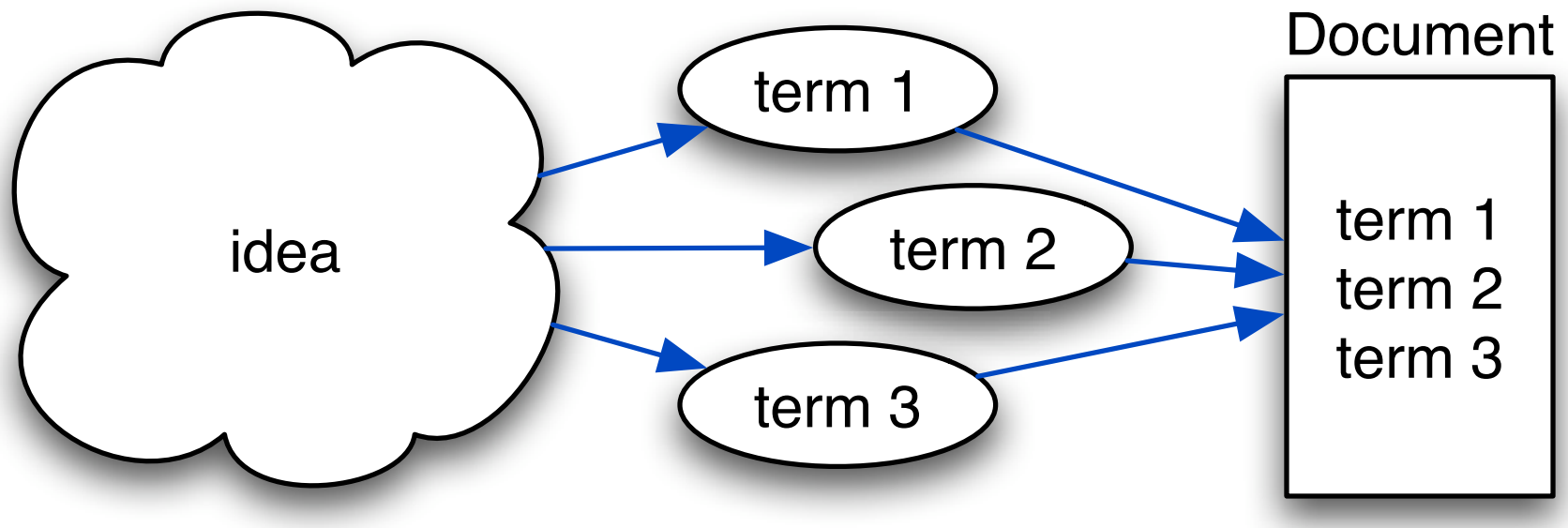
- Topic models for Documents
- Latent Dirichlet Allocation
- Document retrieval with topic models
- Varying the Dirichlet parameter

Topic-based Document Models

Modeling Documents

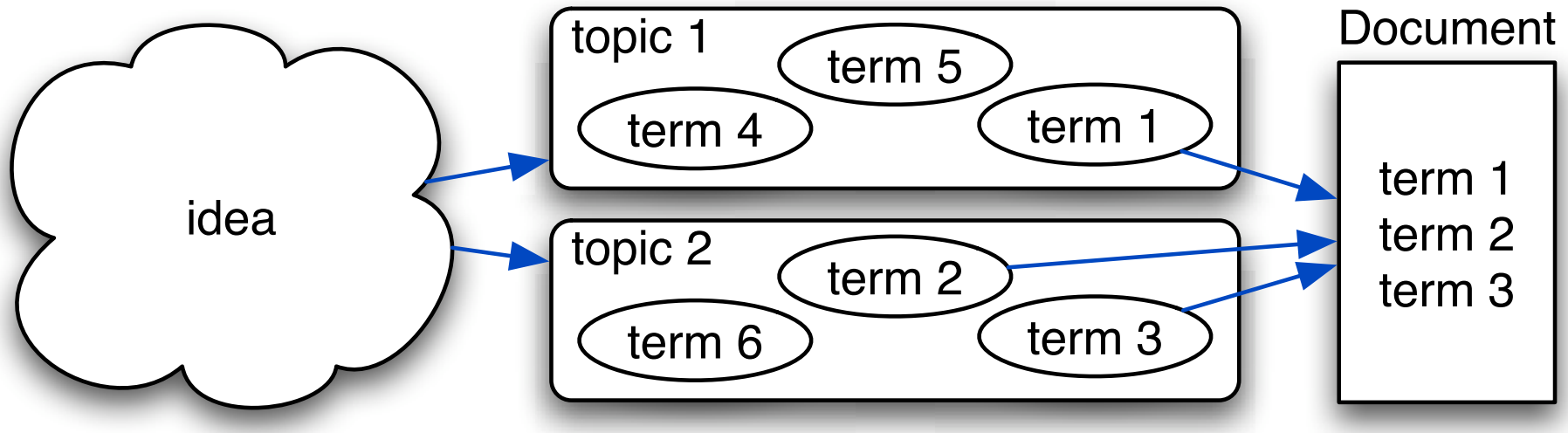
- Documents are a sequence of terms.
- To understand how to model documents, we must examine the purpose of documents.
- Documents are constructed for communication of knowledge. The knowledge exists in the author's mind, the author uses a process to convert this knowledge into a sequence of words.
- By modeling documents, we are attempting to discover the knowledge to text conversion process and hence discover the knowledge that the document is intended to portray.

Naive Document Creation Model



e.g. “baby, clothes”, no “infant” or “suits”

Latent Topic Document Creation Model



Topics are compared, rather than terms.

Latent Dirichlet Allocation

Shape of the Dirichlet Distribution

- The Dirichlet parameter controls the shape of the distribution and hence the likelihood of a topic being selected.

- Each topic has its own parameter, a greater value leads implies that the topic is more probable.

- The exchangeable Dirichlet distribution requires all parameters to be equal, leading to a set of topics having the same likelihood for all topics. Small values lead to only a few topics allocated to each document. This distribution is used in LDA

- The uniform Dirichlet distribution requires that all parameters be set to 1, implying that any combination of topics is equally likely. It has been shown that using a uniform Dirichlet distribution is equivalent to using Probabilistic Latent Semantic Analysis.

Document retrieval with topic models

Topics from Thesaurus

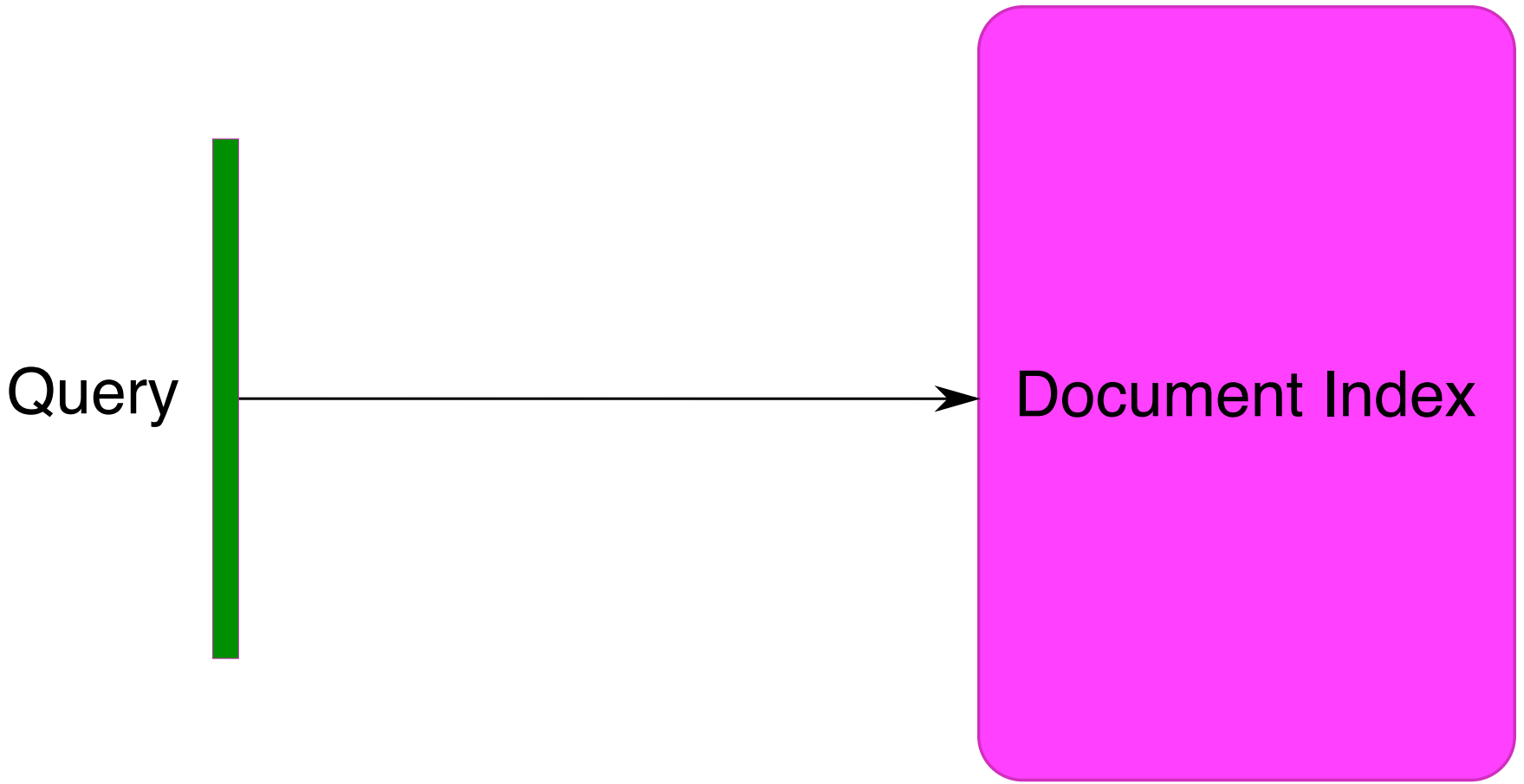
- Topics show us the relationships between terms. A thesaurus is used to look up relationships between terms. By using a thesaurus for query expansion, we are able to apply the expanded query to a term frequency index.

- The thesaurus contains the values:

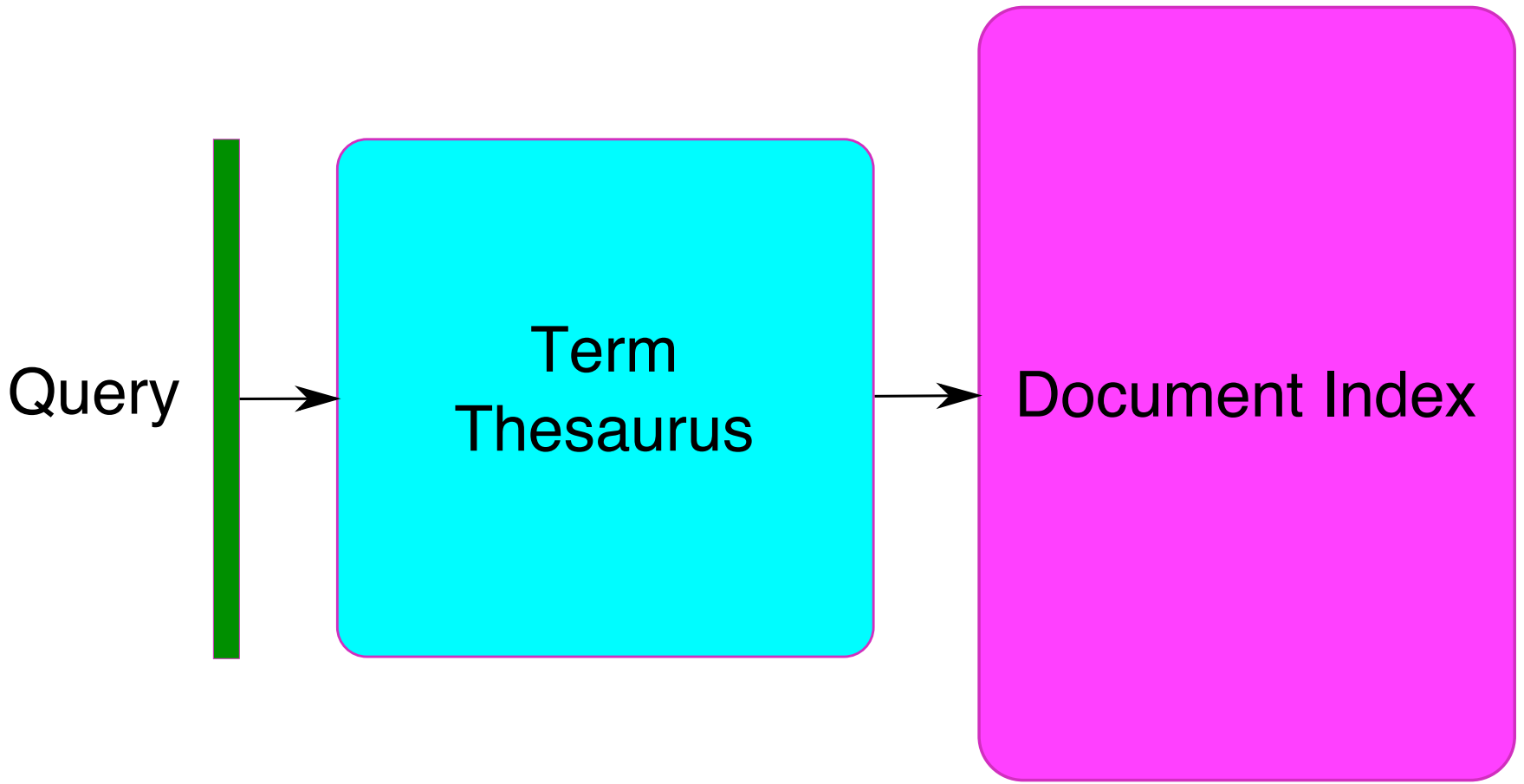
$$
\begin{aligned}
P(t_x|t_y,\alpha) &= \sum_i P(t_x|z_i,\alpha)P(z_i|t_y,\alpha) \\
&= \frac{\sum_i P(t_x|z_i,\alpha)P(t_y|z_i,\alpha)\alpha_i}{\sum_j P(t_y|z_j,\alpha)\alpha_j}
\end{aligned}
$$

- Given an exchangeable Dirichlet distribution, the alphas cancel each other.

Using the thesaurus



Using the thesaurus



Varying the Dirchlet Parameter

How important is the choice of alpha?

- Using an exchangeable Dirichlet distribution, we use the same parameter for every topic, meaning there is only one parameter.

- By varying the alpha parameter, we vary the number of topics that are associated to each term (a high alpha lead to many topics being associated to each term, a low value leads to only a few topics being associated to each term).

- In our experiments, we will examine the effect that fitting alpha has on document retrieval performance, and find if it is necessary to perform the fitting.

- Data set: TREC Disk 2 (Associated Press, Financial Review, Wall Street Journal, Ziff Publishing), with queries 51-200 (used in TREC 1,2,3). The term frequency index was queried using BM25.

Experiment 1: Topic and term inclusion

- How does the exclusion or inclusion of terms and topics affect average precision?

- We built thesauruses for 1) all terms and 100 topics, 2) 100 topics and terms that appear in at least 50 documents, and 3) 300 topics and terms that appear in at least 50 documents.

- The set of 150 queries were used on each of the thesauruses with each document set. Using the two-sided Wilcoxon Signed Rank test, we found no significant difference in each of the three configurations.

- The remainder of the experiments are performed using 300 topics and terms that appear in at least 50 documents.

Experiment 2: Storage and computation times

unfitted alpha

	alpha	Storage	Build time
AP	1	87 MB	57 min
FR	1	29 MB	17 min
WSJ	1	90 MB	61 min
ZIFF	1	42 MB	32 min

Experiment 2: Storage and computation times

unfitted alpha

	alpha	Storage	Build time
AP	1	87 MB	57 min
FR	1	29 MB	17 min
WSJ	1	90 MB	61 min
ZIFF	1	42 MB	32 min

fitted alpha

	alpha	Storage	Build time
AP	0.020	112 MB	2 days
FR	0.021	30 MB	1 day
WSJ	0.022	107 MB	2.5 days
ZIFF	0.033	52 MB	2 days

Experiment 3: Comparison of retrieval precision

- By using a thesaurus, we obtain two parameters (expansion mix and expansion size) that must be set before performing retrieval that will affect the precision.

- We examined expansion mix sizes of 0 to 1 in 0.1 intervals (where the value ranges from 0 to 1), and we examined expansion sized of 10, 20, 50, 100, 200, 500 and 1000 terms (where the expansion size must be greater than or equal to 0). We compares a fitted LDA system to an unfitted LDA system with the same configurations.

- The results showed no significant increase in precision of fitted LDA over unfitted LDA on FR for all expansion parameters, significant increase on ZIFF with some parameters but at the same time not significant over BM25, and a significant increase over both unfitted LDA and BM25 for mix = 0.9 in AP and WSJ. Unfortunately, the best precision was obtained when mix = 0.7, 0.8.

Experiment 3: Best mix, expansion combination

Data	Method	Mix	Expn	MAP	Prec10	MRR
AP	Fitted LDA	0.8	200	0.284	0.380	0.562
	Unfitted LDA	0.8	200	0.282	0.382	0.563
	BM25	NA	NA	0.271	0.355	0.537
FR	Fitted LDA	0.8	1000	0.233	0.141	0.368
	Unfitted LDA	0.8	1000	0.234	0.141	0.367
	BM25	NA	NA	0.197	0.117	0.314

Experiment 3: Best mix, expansion combination

Data	Method	Mix	Expn	MAP	Prec10	MRR
AP	Fitted LDA	0.8	200	0.284	0.380	0.562
	Unfitted LDA	0.8	200	0.282	0.382	0.563
	BM25	NA	NA	0.271	0.355	0.537
FR	Fitted LDA	0.8	1000	0.233	0.141	0.368
	Unfitted LDA	0.8	1000	0.234	0.141	0.367
	BM25	NA	NA	0.197	0.117	0.314

Experiment 3: Best mix, expansion combination

Data	Method	Mix	Expn	MAP	Prec10	MRR
AP	Fitted LDA	0.8	200	0.284	0.380	0.562
	Unfitted LDA	0.8	200	0.282	0.382	0.563
	BM25	NA	NA	0.271	0.355	0.537
FR	Fitted LDA	0.8	1000	0.233	0.141	0.368
	Unfitted LDA	0.8	1000	0.234	0.141	0.367
	BM25	NA	NA	0.197	0.117	0.314

Experiment 3: Best mix, expansion combination

Data	Method	Mix	Expn	MAP	Prec10	MRR
WSJ	Fitted LDA	0.8	500	0.270	0.370	0.625
	Unfitted LDA	0.8	500	0.269	0.369	0.625
	BM25	NA	NA	0.257	0.353	0.572
ZIFF	Fitted LDA	0.7	1000	0.280	0.169	0.427
	Unfitted LDA	0.8	500	0.280	0.165	0.460
	BM25	NA	NA	0.269	0.158	0.415

Experiment 3: Best mix, expansion combination

Data	Method	Mix	Expn	MAP	Prec10	MRR
WSJ	Fitted LDA	0.8	500	0.270	0.370	0.625
	Unfitted LDA	0.8	500	0.269	0.369	0.625
	BM25	NA	NA	0.257	0.353	0.572
ZIFF	Fitted LDA	0.7	1000	0.280	0.169	0.427
	Unfitted LDA	0.8	500	0.280	0.165	0.460
	BM25	NA	NA	0.269	0.158	0.415

Experiment 3: Best mix, expansion combination

Data	Method	Mix	Expn	MAP	Prec10	MRR
WSJ	Fitted LDA	0.8	500	0.270	0.370	0.625
	Unfitted LDA	0.8	500	0.269	0.369	0.625
	BM25	NA	NA	0.257	0.353	0.572
ZIFF	Fitted LDA	0.7	1000	0.280	0.169	0.427
	Unfitted LDA	0.8	500	0.280	0.165	0.460
	BM25	NA	NA	0.269	0.158	0.415

Conclusions

- Latent Dirichlet Allocation (LDA) allocates topics using a Dirichlet distribution
- The Dirichlet distribution has a parameter that must be fitted or estimated.
- We compared the effectiveness of Information retrieval using LDA with both a fitted and estimated Dirichlet parameter.
- We found that LDA using a fitted parameter did not provide a significant increase in precision when operating in peak settings.
- By not fitting the LDA document models, we obtain a 50 to 90 times gain in computational efficiency. Therefore fitting the LDA document models does not provide any benefit for the Information retrieval task.