Evaluating Topic Models for Information Retrieval

Xing Yi and James Allan
Center for Intelligent Information Retrieval, Department of Computer Science
University of Massachusetts, Amherst, MA 01003-4610, USA
{yixing,allan}@cs.umass.edu

ABSTRACT

We explore the utility of different types of topic models, both probabilistic and not, for retrieval purposes. We show that: (1) topic models are effective for document smoothing; (2) more elaborate topic models that capture topic dependencies provide no additional gains; (3) smoothing documents by using their similar documents is as effective as smoothing them by using topic models; (4) topics discovered on the whole corpus are too coarse-grained to be useful for query expansion. Experiments to measure topic models' ability to predict held-out likelihood confirm past results on small corpora, but suggest that simple approaches to topic model are better for large corpora.

Categories and Subject Descriptors: H.3.3 [Information Storage & Retrieval]: Information Search & Retrieval

General Terms: Algorithms, Experimentation

Keywords: Topic Model, Retrieval, Evaluation

1. INTRODUCTION

Topic models are a very popular approach for representing the content of documents. A document is assumed to draw its vocabulary from one or more topics. Topics are represented as probability distributions over the vocabulary, where differing topics give different words high probabilities. We infer a set of topics which can be used to describe the contents of a collection. The high probability topics and words within them can be viewed as a loose description of the collection, with better topic models providing better descriptions. A natural question is whether these topics are useful to help retrieve documents on the same topic as a query – intuitively relevant documents have topic distributions that are likely to have generated the set of words associated with the query[1, 2]. In fact, early research on topic models suggested that they might be used for information retrieval (IR)[2], but it was not until recently that they were successfully applied to large-scale and realistic collections [6]. Our goal in this study is to explore the utility of different topic models for retrieval purposes.

Different topic models specify different document generative procedures, which can lead to very different topics. We use a selection of representative machine-learning topic models: the Mixture of Unigrams (MU) [1], Latent Dirichlet Allocation (LDA)[1, 6] and Pachinko Allocation Model (PAM) [4]. We also construct a special kind of topic model based on Relevance Modeling (RM) [3], which treats each document as the representative of its own topic.

CIKM'08, October 26–30, 2008, Napa Valley, California, USA.
ACM 978-1-59593-991-3/08/10.

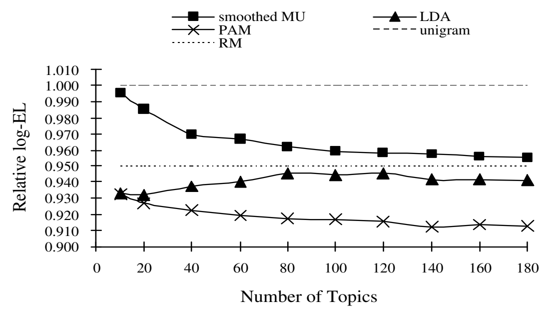

Figure 1: Likelihood comparison on the NIPS dataset. $l_{unigram}$ is -250187. rl_{RM} is 0.9497.

2. DOCUMENT MODELING

We first follow a standard machine learning evaluation approach to compare the document modeling performance of different topic models[1]. We calculate the log Empirical Likelihoods (log-ELs) of generating a held-out test data from different models[4], then further divide the log-ELs by that from the unigram model to obtain relative log-ELs – rl_{TM} for better comparison. Thus, $rl_{unigram}$ is always 1.0 even using different sized test data and low scores indicate better representation of the corpus.

We first use a small NIPS abstract dataset [4], with 1647 abstracts from NIPS proceedings. Following [4], we split the dataset into two subsets with 75% and 25% of the data respectively. We train the models with the larger set and calculate likelihood for the smaller set. We use 50 super-topics for PAM, with the number of sub-topics varying from 10 to 180. Figure 1 shows how the relative log-ELs change with different number of topics and different models. Our results are consistent with previous results[4]: PAM and LDA always perform better than simple models (MU and RM); PAM performs the best. By viewing each document as a topic, RM performs better than MU, and close to LDA. We then use large scale corpora for comparison which has not been done before to the best of our knowledge. We utilize five large TREC corpora used in previous research [5, 6]. They are AP, FT, SJMN, LA, WSJ, with between 90 and 242K documents and roughly 100 queries per corpus. For each of the five corpora, we split it into halves: training different models with one half and calculating likelihood for

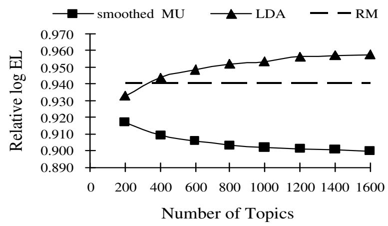

Figure 2: Likelihood comparison with different number of topics on AP. $l_{unigram}$ **is** -2.689×10^8**,** rl_{RM} **is** 0.9408.

	FT	WSJ	SJMN	LA
QL	0.2614	0.2646	0.1612	0.2275
BT-CBQE	0.2628	0.2668	0.1703	0.2228
BT-LBQE	0.2700	0.2696	0.1631	0.2162
BT-PBQE	0.2589	0.2699	0.1637	0.2231
RM-1	0.2783	0.3059	0.1804	0.2488
CBQE	0.2634	0.2628	0.1710	0.2206
LBQE	0.2663	0.2701	0.1656	0.2194
PBQE	0.2607	0.2666	0.1666	0.2151
RM	**0.3006**	**0.3264**	**0.2116**	**0.2605**
CBDM	0.2738	0.2738	0.1802	0.2298
BT-LBDM	0.2681	0.2662	0.1771	0.2330
BT-PBDM	0.2675	0.2738	0.1715	0.2207
RMDE-1	**0.2836**	0.2793	0.1774	0.2457
MBDM	0.2718	0.2771	0.1842	0.2316
LBDM	0.2787	0.2819	**0.1989**	**0.2499**
PBDM	0.2823	0.2815	0.1908	0.2382
RMDE	0.2811	**0.2841**	0.1784	0.2436

Table 1: MAPs of different methods on the testing corpora. Bold font shows the 1st and 2nd best results.

the other half. PAM has not been used in this comparison because its training is too expensive. Figure 2 shows the results on AP test data. It was surprising that the results on large corpora do not mimic those on small corpora: LDA has higher rl than smoothed MU. We speculate that smoothed MU performs better than LDA when using relatively small number of topics for modeling distributions of large corpora.

3. DOCUMENT RETRIEVAL

There are two obvious approaches to including topic models in IR. The first is the document modeling approach, where we calculate $P(w|D)$ by smoothing the document with topics from different topic models: when smoothing the document with the highest ranked topic it is in and using the topic from MU, LDA, PAM and RM, we have retrieval methods – CBDM[5], BT-LBDM, BT-PBDM and RMDE-1, respectively; when smoothing the document with a weighted combination of all topics that it contains and using the topics from MU, LDA, PAM and RM, we have MBDM, LBDM[6], PBDM and RMDE, respectively. The second is the query expansion approach, where we topics similarly add words to the query and run the revised query: when only using the most similar topic from MU, LDA, PAM and RM, we have retrieval methods – BT-CBQE, BT-LBQE, BT-PBQE and RM-1, respectively; when using all the topics from MU, LDA, PAM and RM, we have CBQE, LBQE, PBQE and RM, respectively.

Five TREC corpora and queries are used again for evaluating different topic model based retrieval methods. The same data had been used for evaluating CBDM and LBDM[5, 6]. We use the full AP corpus for training and the other four corpora (FT, SJMN, LA, WSJ) for testing. The number of topics for MU and LDA is tuned to be 2000 and 800 respectively. For PAM, we use 800 sub-topics and 100 super-topics. We also include a simple language modeling retrieval baseline – QL, which does not use topic models and only use Dirichlet smoothing.

Table 1 show the best retrieval results. Our results of CBDM and LBDM are only slightly different from the original results [5, 6] due to small differences in the implementations. We have the following observations: (1) Using topic models for document smoothing can improve IR performance of the typical smoothing technique; complicated topic models like LDA and PAM have some benefits: LBDM and PBDM achieve higher MAPs than CBDM. (2) The document expansion approach RMDE, which borrows idea from RM to do document smoothing and does not actually identify topics in the collection, performs better than CBDM, and sometimes similar to LBDM. (3) PBDM performs similar to LBDM although PAM is more complicated topic modeling approach than LDA. (4) When topic models are trained with the whole corpus, topic models for documents smoothing always outperforms for query expansion.

However, the RM approach, the one that does not actually identify topics in the collection, outperforms all topic modeling approaches consistently. We must conclude that these topic modeling approaches used in these ways are not appropriate for document retrieval. We speculate that the coarse-grained information in full-blown topics is more confusing than useful for this task. We do note that topic model based retrieval methods following the document smoothing approach can be combined with RM [5, 6] to further improve the IR performance although the improvement is very small.

4. CONCLUSIONS

In this paper, we have explored the utility of different types of topic models for IR. Experimental results show that topics trained on the full corpus are more useful for document smoothing than for query expansion. Applying complicated models like LDA for document smoothing can improve IR performance, but more powerful model like PAM does not necessary provide further benefits. RM outperforms all these topic modeling approaches consistently in most cases: one possible reason is that topics discovered by topic models in large scale corpus are not as fine-grained as the query-specific topic calculated by RM.

5. ACKNOWLEDGMENTS

This work was supported in part by the CIIR and in part by the Defense Advanced Research Projects Agency (DARPA) under contract number HR0011-06-C-0023. Any opinions, findings and conclusions or recommendations expressed in this material are the authors' and do not necessarily reflect those of the sponsor.

6. REFERENCES

[1] D. Blei, A. Ng, and M. Jordan. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[2] T. Hofmann. Probabilistic latent semantic indexing. In *Proc. of ACM SIGIR*, pages 50–57, 1999.

[3] V. Lavrenko and W. B. Croft. Relevance-based language models. In *Proc. of ACM SIGIR*, pages 120–127, 2001.

[4] W. Li and A. McCallum. Pachinko Allocation: DAG-structured mixture models of topic correlations. In *Proc. of the 23rd ICML*, pages 577–584, Pittsburgh, PA, 2006.

[5] X. Liu and W. B. Croft. Cluster-based retrieval using language models. In *Proc. of ACM SIGIR*, pages 186–193, 2004.

[6] X. Wei and W. B. Croft. LDA-based document models for ad-hoc retrieval. In *Proc. of ACM SIGIR*, pages 178–185, 2006.