A comparative study of utilizing LDA (Latent Dirichlet Alloction) topic model for information retrieval

JIE SONG, University of Massachusetts, Amherst

1 INTRODUCTION

Topic model is a very popular document representing model nowadays. In the topic model, topics are the vocabulary word's distribution, which is represented as the words' probabilities in the topic. However, the answer for the question if the topic model can improve IR is still not so clear. So, the objective of this experiment is try to prove that the topic model can improve IR. And here we use the LDA topic model to do the experiment. LDA model is a very popular probabilistic topic model, in which each document has several topics and each topic have different distribution in different documents. LDA model is a probabilistic model with a corresponding generative process, in which θ-the topic distribution in each document and φ-the word distribution in each topic is the most two important parameters to estimate. In this experiment, I use the Gibbs sampling method to estimate θ and φ by MALLET [1]. In this experiment, three baseline models – QL (with Dirichlet smoothing) model, DFR-PL2 and DFR-DL2 are used to compare with LBDM model. Two data collections -Trec123 and Robust04, including the query sets and relevance judgments, were used in the experiment. Each parameter in the four different models was estimated by statistic cross validation machine learning method. The performance of each model were evaluated by the mean of average precision(AP), nDCG@10 and ERR@10 metrics, which is compared to the experiment conducted by Xin and W.Bruce [2].

2 EXPERIMENTAL AND COMPUTATIONAL DETAILS

2.1 Probabilistic Model Representation

Four models were built to conduct the experiment:
1. QL with Dirichlet smoothing:

$$p(w|D) = \frac{N_D}{N_D+\mu} p_{ML}(w|D) + \frac{\mu}{N_D+\mu} p(w|\mathcal{C}).$$

2. DFR-PL2

$$weight(w|d) = \frac{1}{tfn+1}(tfn \cdot \log_2 \frac{tfn}{\lambda} + (\lambda + \frac{1}{12 \cdot tfn} - tfn) \cdot \log_2 e + 0.5 \cdot \log_2(2\pi \cdot tfn))$$

where $tfn = tf \cdot \log_2(1 + c \cdot \frac{avg_l}{l(d)})$

3. DFR-DL2

$$weight(w|d) = F \cdot D(\varphi, p) + 0.5 \cdot \log_2(2\pi \cdot tfn(1-\varphi))$$

where $\varphi = \frac{tfn}{F}, p = \frac{1}{n}, D(\varphi, p) = \varphi \cdot \log_2 \frac{\varphi}{p} + (1-\varphi) \cdot \log_2((1-\varphi)(1-p)),$

$$tfn = tf \cdot \log_2(1 + c \cdot \frac{avg_l}{l(d)})$$

4. LBDM

$$P(w\mid D)=\lambda\left(\frac{N_d}{N_d+\mu}P'(w\mid D)+\left(1-\frac{N_d}{N_d+\mu}\right)P'(w\mid coll)\right)$$
$$+(1-\lambda)\left(\sum_{t=1}^{K}\frac{n_{-i,j}^{(w_i)}+\beta_{w_i}}{\sum_{v=1}^{V}(n_{-i,j}^{(v)}+\beta_v)}\times\frac{n_{-i,j}^{(d_i)}+\alpha_{z_i}}{\sum_{t=1}^{T}(n_{-i,t}^{(d_i)}+\alpha_t)}\right) \quad (9)$$

2.2 Parameter Estimate

2.2.1 LDA model. Cross validation was conducted to estimate the parameters in each model. Actually, for the LBDM model, there are 6 parameters in total: α-Dirichlet distribution for each document, β-Dirichlet distribution for each topic, K-number of topics, n-iteration numbers for Gibbs sampling, μ and λ. The common setting for α and β is: α = 50/K, β = 0.01 [2,3], which we also use this setting in this experiment. And in Xin and W.Bruce's [2] experiment, they have proved that 30 iterations for the Gibbs sampling is enough for the LDA model, considering the running time of LDA model program, here I also use this setting. And for the K setting, since the running time of each new LDA model building with 30 iterations' Gibbs sampling is too long, I did not use cross validation method to estimate the best K setting. Instead, I compare different K values in three LDA model running. Table 1 is the retrieval results for different K value.

Table 1. Retrieval results on AP, nDCG and ERR with different number of topic K

TREC123				ROBUST04			
K	mean AP	nDCG@10	ERR@10	K	mean AP	nDCG@10	ERR@10
200.0000	0.2272	0.5045	0.4251	200.0000	0.2557	0.5033	0.4982
400.0000	0.2273	0.5045	0.4250	400.0000	0.2556	0.5024	0.4982
600.0000	0.2273	0.5045	0.4250	600.0000	0.2556	0.5025	0.4983

From the results, we can see that there are no significant differences among the setting on K from 200-600. Considering that the bigger K value the running time is longer, I conduct my experiment on LDA model by setting K value to 200. For the smoothing parameters μ and λ, cross validation was used to estimate the best setting for each data collections (shown in Table.2).

Table 2. Retrieval results on AP, nDCG and ERR with different parameters for each model

Parameter Estimate	TREC123			ROBUST04		
	mean AP	nDCG@10	ERR@10	mean AP	nDCG@10	ERR@10
QL-μ	2000.0	2000.0	3500.0	1000.0	1000.0	3500.0
DL2-c	5.0	4.0	2.0	9.0	5.0	5.0
PL2-c	6.0	6.0	8.0	9.0	6.0	6.0
LDA-μ	2000.0	2000.0	2000.0	1000	1000	3000
LDA-λ	0.3	0.3	0.3	1.0	1.0	1.0

2.2.2 Baseline model. The best estimate setting for the μ of QL model and c for DFR-PL2 and DFR-DL2 is validated by cross validation method. The results are shown in Table 2. Different data collections have different parameter settings for a same model.

3 RESULTS AND DISCUSSION

3.1 Retrieval Experiments

The retrieval results were saved in trec123.txt and robust04.txt files, and the data is presented in Table 3. Based on the result, for Trec123, statistically significant improvement of LBDM over QL on mean AP metric was overserved. Also, LBDM significantly improve the IR results over DFR-PL2 and DFR-DL2 by ERR metric. No significant improvement was observed according to nDCG metric for Trec123. While for Robust04 data collection, LBDM significantly improve the retrieval results over QL on both mean AP and nDCG metric. However, it significantly decreases the results over DFR-PL2 and DFR-DL2 by nDCG metric.

Table 3. Retrieval and t test results for comparing LBDM with QL, DFR-PL2 and DFR-DL2

TREC123							
	QL	DFR-PL2	DFR-DL2	LBDM	p-value over QL	p-value over DFR-PL2	p-value over DFR-DL2
mean AP	0.2301	0.2280	0.2282	0.2302	0.0495**+	0.2241	0.2597
nDCG@10	0.5129	0.5072	0.5070	0.5138	0.1756	0.3307	0.3120
ERR@10	0.4380	0.4252	0.4254	0.4382	0.2817	0.0492**+	0.0529*+
Robust04							
	QL	DFR-PL2	DFR-DL2	LBDM	p-value over QL	p-value over DFR-PL2	p-value over DFR-DL2
mean AP	0.2526	0.2571	0.2567	0.2556	0.0306**+	0.3422	0.4933
nDCG@10	0.4955	0.5117	0.5118	0.5032	0.0901*+	0.0140**-	0.0142**-
ERR@10	0.5026	0.5025	0.5018	0.4981	0.4399	0.3319	0.4222

**: significant differ at 95% confidence, *: significant differ at 90% confidence
+: improve, -: decrease

3.2 Comparisons and Discussion

In Xin and W.Bruce's [2] and Xing and James's [4] experiments, they use 5 other data collections (AP,FT,WSJ,SJMN and LA) and evaluated the performance based on mean AP metric. They found that LBDM significantly improve the IR result over QL model, which is exactly consistent with our experiment result. While compared to DFR-PL2 and DFR-DL2, LBDM significantly improve the results for Trec123 but significantly decrease the result for Robust. From this result, we can conclude that different data collections (or corpus) will achieve different IR results even by the same retrieval technique, which means we do the information retrieval, we should consider the corpus effect on different on different retrieval models.

Another import fact that would affect the IR result is the quality of the queries. As we mentioned above, DFR model is significant better than LBDM for robust04; while for trec123, LBDM is significant better than DFR. Based on this finding, when we looked into each query for trec123 and robust04, an interesting finding is that over 80% of trec123 queries have more than 100 relevant documents; while over 70% of robust04 queries have less than 100 relevant documents (data not show here, saved in analysis.txt). This means the quality of the queries for the two data collections is totally different, which may explain why we got the contrary results on comparing LBDM model with DFR models for the two data collections. As we know, the DFR is a probabilistic model based on the frequency of words, while the LDA model is one based on the latent meaning of words. The fact that the query has more relevant documents in the data collections probably indicate that the query may be a common word in that corpus. When we use common words or unique words to do the information retrieval, the results should be definitely different. From our results, it showed that LBDM works better than DFR or QL when using command words as queries, while when using less common words as queries, LBDM still works better than QL but not better than DFR model. To further investigate the effect of the quality of queries on IR search results, it is better to combine LDA model with RM language model to do the information retrieval. RM1 model will do the query expansion to decrease the effect of queries' quality on the search. Actually, in Xin and W.Bruce's [2] paper, they did combine LBDM with RM, and the results showed that moderate improvements are obtained. I also did this combination for LBDM and RM1, in which I picked 20 queries whose LBDM results were significant worse than DFR-PL2 and DFR-DL2. However, the results are not as good as the paper mentioned (the result is saved in LDA-RM1.txt).

Last but not the least, although LBDM did a good job in improving IR result, one thing need to be pointed out is that the running time for LBDM is really much longer then QL, DFR and even RM. Further study should be focused on how to improve the efficiency of LBDM model.

4 CONCLUSIONS

In summary, we have performed a reproducibility experiment on comparing LDA model and QL and DFR models. The results are mostly consistent with other's work – LBDM model can significantly improve the IR results by comparing with other models, except that DFR models work better than LBDM for Robust04 with given queries. Two factors determine IR search results by a same model: corpus and queries. The combination of topic model and language model will obtain better improvement on IR. Further study should be conducted to confirm the conclusion and the work efficiency of LBDM model should be improved.

REFERENCES

[1] http://mallet.cs.umass.edu/topics.php

[2] Xing Wei and W. Bruce Croft. 2006. *LDA-Based Document Models for Ad-hoc Retrieval. ACM 1-58113-000-0/00/0004*

[3] M Steyvers, T Griffiths. 2007. *Probabilistic topic models.* Handbook of latent semantic analysis 427 (7), 424-440

[4] Xing Yi and James Allan. 2009. *A comparative study of utilizing topic models for information retrieval.* M. Boughanem et al. (Eds.): ECIR 2009, LNCS 5478, pp. 29–41