A Study of Document Expansion using Translation Models and Dimensionality Reduction Methods

Saeid Balaneshin-kordan
Department of Computer Science
Wayne State University
Detroit, Michigan 48202
saeid.balaneshinkordan@wayne.edu

Alexander Kotov
Department of Computer Science
Wayne State University
Detroit, Michigan 48202
kotov@wayne.edu

ABSTRACT

Over a decade of research on document expansion methods resulted in several independent avenues, including smoothing methods, translation models, and dimensionality reduction techniques, such as matrix decompositions and topic models. Although these research avenues have been individually explored in many previous studies, there is still a lack of understanding of how state-of-the-art methods for each of these directions compare with each other in terms of retrieval accuracy. This paper attempts to fill in this void by reporting the results of an empirical comparison of document expansion methods using translation models estimated based on word co-occurrence and cosine similarity between low-dimensional word embeddings, Latent Dirichlet Allocation (LDA) and Non-negative Matrix Factorization (NMF), on standard TREC collections. Experimental results indicate that LDA-based document expansion consistently outperforms both types of translation models and NMF according to all evaluation metrics for all and difficult queries, which is closely followed by translation model using word embeddings.

CCS Concepts

•Information systems → Document collection models;

Keywords

Document Expansion; Translation Models; Latent Dirichlet Allocation; Non-negative Matrix Factorization; Word Embeddings

1. INTRODUCTION

Vocabulary mismatch is one of the fundamental problems in Information Retrieval (IR), which in the context of language modeling approaches, has been traditionally addressed through expansion of document or query language models (LM) with semantically related terms. However, finding such terms for a particular document or query is a challenging task. While query expansion methods are applied on-line and typically identify such terms in external resources or top-retrieved documents (i.e. using local analysis [19]), document expansion can be done off-line and attempts to identify semantic structure in a collection (i.e. using global analysis [19]).

Depending on their theoretical foundation, the proposed approaches to identifying such structure resulted in several independent research avenues. Statistical translation models [1] quantify the strength of semantic relationship between pairs of words. Translation models estimated using mutual information [7] or term co-occurrence [8] have been shown to improve the performance of language modeling based retrieval models. In addition to these methods, the utility of word embeddings [16] has also been evaluated for estimation of translation models [22].

Dimensionality reduction techniques, such as Latent Dirichlet Allocation (LDA) [2], Probabilistic Latent Semantic Indexing (pLSI) [6], and Non-Negative Matrix Factorization (NMF) [13, 20], approximate document collection using its lower dimensional representations. In particular, topic models (LDA and pLSI) estimate the parameters of a probabilistic generative process, while NMF approximates sparse high-dimensional document-term space with dense low-dimensional subspaces. Although LDA has been shown to be effective for ad hoc IR by several previous studies [18, 21], NMF has been primarily studied in the context of text mining [11], and its application to IR requires further investigation.

Although these research avenues have been individually explored in previous studies, there is a lack of understanding of how state-of-the-art methods for them compare with each other in terms of retrieval accuracy. This paper attempts to fill this void by reporting the results of an empirical comparison of document expansion methods using LDA, NMF, and translation models estimated based on word co-occurrence and cosine similarity between word embeddings on newswire and Web TREC collections.

2. RELATED WORK

Translation models [1] have been extensively studied in the context of IR. Karimzadehgan and Zhai [8] have shown that the translation model derived using axiomatic framework and estimated based on term co-occurrences outperforms the translation model estimated based on mutual information [7]. In [22, 4], word embeddings were integrated into language modeling based retrieval models. Zuccon et al. [22] have compared estimation of translation model using

ICTIR '16, September 12-16, 2016, Newark, DE, USA

DOI: http://dx.doi.org/10.1145/2970398.2970439

word embeddings and mutual information [8]. The translation model proposed in [4] is a generalized version of the model in [22], which also accounts for the fitness of query terms in the context of a document.

Topic models (LDA and pLSI) consider documents as finite mixtures over an underlying set of latent topics inferred from correlations between words. LDA has been proposed as an improvement of pLSI, in which the mixtures of topics for documents are assumed to be drawn from the same Dirichlet prior. This modification makes LDA more robust to overfitting than pLSI and other more restrictive models, such as the mixture of unigrams. Nevertheless, Lu et al. [14] found out that document expansion methods based on pLSI and LDA have comparable retrieval accuracy on different collections. They also observed that topic models can hurt retrieval performance of document expansion if they are not properly applied and optimized. Masada et al. [15] compared pLSI and LDA for classification tasks in terms of computation time and found out that LDA does not offer any significant improvement over pLSI in terms of F-measure, while training LDA requires more time. Ding et al. [3] found out that NMF with I-divergence objective function and pLSI both have very close accuracy, entropy, purity, Rand index when used for document clustering, while Gaussier and Goutte [5] have shown that NMF and pLSI are equivalent in the sense that both optimize the same objective function.

3. METHODS

We used the query likelihood retrieval model [17] with Dirichlet prior smoothing [12] (`QL-DIR`) in conjunction with all document expansion methods. Given query q, this model calculates the retrieval score of each document d as:

$$p(d|q) = \prod_{w \in q} \left(\frac{c(w;d) + \mu p(w|C)}{|d| + \mu} \right) \quad (1)$$

where $c(w;d)$ is the count of word w in document d, $p(w|C)$ is the collection language model, $|d|$ is the length of document d and μ is the Dirichlet prior. Basic language modeling approaches, such as `QL-DIR`, are based on exact matching of query terms. Since queries are typically short and relevant documents may use different vocabulary, such models often suffer from vocabulary mismatch. Incorporating document expansion language model (LM) constructed using the methods discussed below into the original document LM can potentially address this problem. In this case, (1) can be rewritten as:

$$p(d|q) = \prod_{w \in q} \left((1-\lambda) \frac{c(w;d) + \mu p(w|C)}{|d| + \mu} + \lambda p_{\text{exp}}(w|d) \right) \quad (2)$$

where $p_{\text{exp}}(w|d)$ is the document expansion LM and λ is the interpolation coefficient.

3.1 Translation Models

Translation models [1] estimate document expansion LMs on a term-by-term basis by "translating" each original document term u according to the translation probabilities $p_{\text{tr}}(w|u)$. Each "translation" constitutes addition of one or several semantically related terms to the original document LM. Therefore, a document expansion LM is constructed using translation model as follows:

$$p_{\text{t}}(w|d) = \sum_{u \in \mathbb{V}} p_{\text{tr}}(w|u) p_{\text{ml}}(u|d) \quad (3)$$

In (3), $p_{\text{ml}}(u|d) = c(u;d)/|d|$ is the original document LM, where $c(u;d)$ is the number of occurrences of word u in document d. $p_{\text{tr}}(w|u)$ in the above equation can be obtained from the number of co-occurrences of word w with word u (i.e., $c(w,u)$) and the number of co-occurrences of word w with other words in the collection vocabulary as follows (`TM-CX`) [8]:

$$p_{\text{tr}}(w|u) = \frac{c(w,u)}{\sum_{v \in \mathbb{V}} c(v,u) + |V|} \quad (4)$$

where $|V|$ is the size of collection vocabulary V.

Another way to approximate the translation model is to utilize word embeddings that are pre-trained for the document collection [16]. In this method, which is denoted by `TM-WE`, semantic similarity between the words in the word embeddings space is calculated based on the cosine similarity of their corresponding word vectors [22]. The translation probability is obtained by normalizing this cosine similarity.

3.2 Latent Dirichlet Allocation

Topic models, such as LDA and its extensions, can also be used to construct document expansion LMs [9, 10, 18] based on the assumption that the words belonging to the same topic are semantically related. LDA considers each document d in the collection as a mixture of multinomials (topics) z drawn from a symmetric Dirichlet prior $\text{Dir}(\alpha)$ with parameter α and models it according to the following generative process:

- for each document d, draw a distribution over topics (i.e., $p_\theta(z|d)$) from $\text{Dir}(\alpha)$
  - for each word position in d, draw a topic z from the distribution $p_\theta(z|d)$
  - draw a word w from the distribution $p_\phi(w|z)$.

where $p_\theta(z|d)$ and $p_\phi(w|z)$ represent the probability distributions of topics in document d and words in topic z, respectively. The vocabulary gap can be eliminated by including the terms in the topics that have a high probability in the topic distribution of a document into the expansion LM for that document. Document expansion LM can be constructed based on the output of topic models as follows:

$$p_{\text{lda}}(w|d) = \sum_{z \in Z} p_\phi(w|z) p_\theta(z|d) \ , \quad (5)$$

where Z is the number of topics.

3.3 Non-negative Matrix Factorization

Similar to LDA, NMF can also be used to discover topics in word-document matrix, $\mathbf{P}$. If the probability of word w given topic z is denoted by $p_{\text{b}}(w|z)$ and the likelihood of topic z given document d is denoted by $p_{\text{e}}(z|d)$, then:

$$p_{\text{nmf}}(w|d) = \sum_{z \in Z} p_{\text{b}}(w|z) p_{\text{e}}(z|d) \quad (6)$$

This equation can be written in matrix form as $\mathbf{P} = \mathbf{P}_{\text{b}} \mathbf{P}_{\text{e}}$, where $\mathbf{P}_{\text{b}}$ and $\mathbf{P}_{\text{e}}$ are non-negative matrices of size $K \times R$ and $R \times M$, respectively, the inner dimension (R) of which can be considered as the number of "topics". It is assumed

that $R < \min(K, M)$, therefore the matrices $\mathbf{P}_\mathrm{b}$ and $\mathbf{P}_\mathrm{e}$ are lower-dimensional factors, the product of which approximates $\mathbf{P}$. The matrices $\mathbf{P}_\mathrm{b}$ and $\mathbf{P}_\mathrm{e}$ are obtained by solving the following optimization problem:

$$\min_{\mathbf{P}_\mathrm{b}, \mathbf{P}_\mathrm{e}} \frac{1}{2} \sum_i \sum_j [\mathbf{P}_{i,j} - (\mathbf{P}_\mathrm{b}\mathbf{P}_\mathrm{e})_{i,j}]^2 \quad (7)$$

In the above problem, $\mathbf{P}_{i,j}$ is the (i,j)-th element of the TF-IDF document-term matrix $\mathbf{P}$ and $(\mathbf{P}_\mathrm{b}\mathbf{P}_\mathrm{e})_{i,j}$ is the (i,j)-th element of the matrix $(\mathbf{P}_\mathrm{b}\mathbf{P}_\mathrm{e})$. Although all the elements of $\mathbf{P}_\mathrm{b}$ and $\mathbf{P}_\mathrm{e}$ are non-negative, they should be normalized to represent probabilities.

4. EXPERIMENTS

Experimental evaluation of document expansion methods using the translation models and dimensionality reduction techniques presented in the previous section was performed using standard TREC collections (TREC7-8, ROBUST04, and GOV) and query sets (100 topics from TREC 2007-2008 Ad Hoc track, 250 topics from TREC 2004 ROBUST track, 225 topics from TREC 2004 Web track).

All retrieval methods were implemented using Indri 5.9 IR toolkit[1]. Experimental collections were pre-processed using INQUERY stoplist and Porter stemmer. Very frequent terms (that occur in more than 15% of documents) and very rare terms (that occur in less than 5 documents) were ignored for estimation of translation models, LDA and NMF. We used the implementation of NMF from scikit-learn 17.1[2]. We used Double Singular Value Decomposition for non-random initialization of factors and Coordinate Descent solver for optimization. The values of LDA hyperparameters α and β for the Dirichlet priors in LDA were set to $1/Z$ (where Z is the number of topics) and 0.01, respectively. Gibbs sampler for posterior inference of LDA parameters was run for 200 iterations, and the convergence threshold for NMF was set to 10^{-5}. Word embeddings were obtained by using word2vec (version 0.1c) tool[3] (by using Skip-gram architecture for its training). The following values for the parameters of the word2vec were used: size of word vectors was set to 100, max skip length between words, threshold for word occurrence and the number of negative examples were set to 5, and the starting learning rate was set to 0.025.

Parameters of translation models (# of translated words) and dimensionality reduction techniques (# of topics for LDA and the inner dimension for NMF), as well as the interpolation coefficient of the original document LM and document expansion LM for all methods were empirically optimized based on the Mean Average Precision (MAP) for each of the datasets separately. These parameters are determined at each step of a three-fold cross-validation by using grid search with step size 0.1 for continuous parameters between 0 and 1 and step size 200 for discrete parameters, like the number of topics.

Figure 1 illustrates how the retrieval accuracy of document expansion based on LDA and NMF changes depending on the number of topics (inner components). As follows from Figure 1, document expansion based on LDA clearly outperforms document expansion based on NMF. Retrieval accuracy of LDA-based document expansion has a pronounced peak at around 1000 topics for TREC7-8 and around 1500 for ROBUST04 and GOV, after which it saturates and slowly decreases. It is evident that NMF, on the other hand, peaks early, when the number of inner components is around 100 for GOV, around 300 for TREC7-8 and around 400 for ROBUST04 and then its effectiveness remarkably deteriorates. We attribute this to an empirical observation that a large number of popular words, which are not useful for retrieval, appear in many NMF topics.

Table 1 summarizes retrieval accuracy of `QL-DIR` and document expansion methods using different types of translation models, LDA and NMF across different collection for both all and difficult queries macro-averaged based on 3-fold cross validation. Difficult queries are defined as the ones for which the average precision of `QL-DIR` is less than 0.05. Several important conclusions can be made based on the results in this table. First, the LDA-based document expansion (`LDA`) achieves the best performance according to all metrics, outperforming both types of translation models and `NMF`. Particularly significant improvement of `LDA` over `TM-CX` (∼9% for all queries and 162% for difficult queries) is achieved on the TREC7-8 dataset. Using the translation model based on word embeddings (`TM-WE`) generally results in smaller yet comparable improvement to `LDA`. `TM-WE` is however much less computationally expensive document expansion method than `LDA`. Second, while `NMF` generally outperformed `TM-CX` in terms of all metrics, it had lower MAP than `TM-WE` on all collections for all queries and on TREC7-8 and GOV for difficult queries. `TM-WE` was particularly more effective than `NMF` for all and difficult queries on TREC7-8. Third, estimating translation model based on cosine similarity between word embedding vectors (`TM-WE`) is consistently more effective than using Conditional Context (`TM-CX`) for all queries, and particularly for difficult ones.

5. CONCLUSION

In this paper, we attempted to fill in the void in theoretical IR literature by performing a comparative study of retrieval effectiveness of document expansion methods based on different types of translation models with the ones based on dimensionality reduction techniques, such as topic models and matrix decomposition, on publicly available collections of different size and type. We found out that, although LDA-based document expansion generally outperforms document expansion methods based on NMF and translation models, its performance is comparable to document expansion using translation model estimated based on word embeddings.

6. REFERENCES

[1] A. Berger and J. Lafferty. Information retrieval as statistical translation. In *Proceedings of the 22nd ACM SIGIR*, pages 222–229, 1999.

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

[3] C. Ding, T. Li, and W. Peng. On the equivalence between non-negative matrix factorization and probabilistic latent semantic indexing. *Computational Statistics & Data Analysis*, 52(8):3913–3927, 2008.

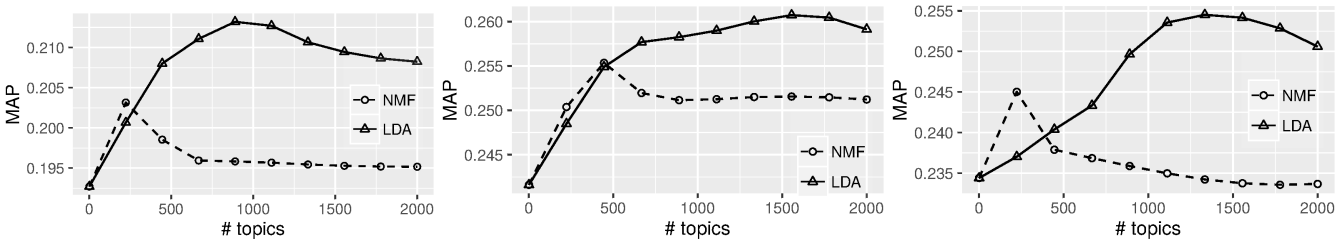
[4] D. Ganguly, D. Roy, M. Mitra, and G. J. Jones. Word embedding based generalized language model for information retrieval. In *Proceedings of the 38th ACM SIGIR*, pages 795–798, 2015.

[5] E. Gaussier and C. Goutte. Relation between pLSA and NMF and implications. In *Proceedings of the 28th ACM SIGIR*, pages 601–602, 2005.

[1] http://www.lemurproject.org/indri.php

[2] http://scikit-learn.org/

[3] http://code.google.com/archive/p/word2vec/



(a) TREC7-8 (b) ROBUST04 (c) GOV

Figure 1: **MAP of NMF and LDA-based document expansion methods for different number of topics.**

Table 1: **Retrieval performance of document expansion methods for all and difficult queries. $*$ and † indicate statistically significant improvement in terms of MAP ($p < 0.05$) using Wilcoxon signed rank test over the QL-DIR and TM-CX, respectively. Relative improvement over QL-DIR and TM-CX is shown in parenthesis.**

Col.	Method	All Queries			Difficult Queries		
		MAP	**NDCG@20**	**P@20**	**MAP**	**NDCG@20**	**P@20**
TREC7-8	QL-DIR	0.1927	0.4142	0.339	0.0205	0.1303	0.1326
	TM-CX	0.1963	0.4149	0.3603	0.0216	0.1346	0.1165
	TM-WE	0.2084 $*$† (+8.1%/+6.2%)	0.4155 (+0.3%/+0.1%)	0.3678 (+8.5%/+2.1%)	0.0434 $*$† (+111.7%/+100.9%)	0.1431 (+9.8%/+6.3%)	0.151 (+13.9%/+29.6%)
	NMF	0.2035 $*$† (+5.6%/+3.7%)	0.4143 (+0.0%/-0.1%)	0.3655 (+7.8%/+1.4%)	0.0229 $*$ (+11.7%/+6.0%)	0.1317 (+1.1%/-2.2%)	0.1174 (-11.5%/+0.8%)
	LDA	**0.2138** $*$† **(+10.9%/+8.9%)**	**0.4312 (+4.1%/+3.9%)**	**0.3745 (+10.5%/+3.9%)**	**0.0565** $*$† **(+175.6%/+161.6%)**	**0.1605 (+23.2%/+19.2%)**	**0.1652 (+24.6%/+41.8%)**
ROBUST04	QL-DIR	0.2416	0.4065	0.3504	0.023	0.0997	0.0962
	TM-CX	0.2543	0.4104	0.3607	0.0353	0.1004	0.0972
	TM-WE	0.2582 $*$† (+6.9%/+1.5%)	0.4191 (+3.1%/+2.1%)	0.3634 (+3.7%/+0.7%)	0.0345 $*$ (+50.0%/-2.3%)	0.1026 (+2.9%/+2.2%)	0.0979 (+1.8%/+0.7%)
	NMF	0.2557 $*$† (+5.8%/+0.6%)	0.4188 (+3.0%/+2.0%)	0.3641 (+3.9%/+0.9%)	0.0348 $*$ (+51.3%/-1.4%)	0.1012 (+1.5%/+0.8%)	0.0983 (+2.2%/+1.1%)
	LDA	**0.2608** $*$† **(+7.9%/+2.6%)**	**0.4197 (+3.2%/+2.3%)**	**0.3629 (+3.6%/+0.6%)**	**0.0353** $*$ **(+53.5%/+0.0%)**	**0.1052 (+5.5%/+4.8%)**	**0.101 (+5.0%/+3.9%)**
GOV	QL-DIR	0.2344	0.3942	0.5141	0.0183	0.1214	0.0353
	TM-CX	0.2449	0.4021	0.5191	0.0216	0.1225	0.0371
	TM-WE	0.2515 $*$† (+7.3%/+2.7%)	0.4101 (+4.0%/+2.0%)	0.5329 (+3.7%/+2.6%)	0.0276 $*$ (+50.8%/+27.8%)	0.1278 (+5.2%/+4.3%)	0.0383 (+8.9%/+3.2%)
	NMF	0.2467 $*$ (+5.2%/+0.7%)	0.4082 (+3.6%/+1.5%)	0.5289 (+2.9%/+1.9%)	0.0232 $*$ (+26.8%/+7.4%)	0.1227 (+1.0%/+0.1%)	0.0379 (+7.4%/+2.2%)
	LDA	**0.2539**$*$† **(+8.3%/+3.7%)**	**0.4131 (+4.8%/+2.7%)**	**0.5365 (+4.4%/+3.3%)**	**0.0296** $*$† **(+61.7%/+37.0%)**	**0.1327 (+9.3%/+8.3%)**	**0.0394 (+11.6%/+6.2%)**

[6] T. Hofmann. Probabilistic latent semantic indexing. In *Proceedings of the 22nd ACM SIGIR*, pages 50–57, 1999.

[7] M. Karimzadehgan and C. Zhai. Estimation of statistical translation models based on mutual information for ad hoc information retrieval. In *Proceedings of the 33rd ACM SIGIR*, pages 323–330, 2010.

[8] M. Karimzadehgan and C. Zhai. Axiomatic analysis of translation language model for information retrieval. In *Proceedings of the 34th ECIR*, pages 268–280. 2012.

[9] A. Kotov, V. Rakesh, E. Agichtein, and C. K. Reddy. Geographical latent variable models for microblog retrieval. In *Proceedings of the 38th ECIR*, pages 635–647, 2015.

[10] A. Kotov, Y. Wang, and E. Agichtein. Leveraging geographical metadata to improve search over social media. In *Proceedings of the 22nd WWW*, pages 151–152, 2013.

[11] D. Kuang, J. Choo, and H. Park. Nonnegative matrix factorization for interactive topic modeling and document clustering. In *Partitional Clustering Algorithms*, pages 215–243. 2015.

[12] J. Lafferty and C. Zhai. Document language models, query models, and risk minimization for information retrieval. In *Proceedings of the 24th ACM SIGIR*, pages 111–119, 2001.

[13] D. D. Lee and H. S. Seung. Algorithms for non-negative matrix factorization. In *Proceedings of NIPS*, pages 556–562, 2001.

[14] Y. Lu, Q. Mei, and C. Zhai. Investigating task performance of probabilistic topic models: an empirical study of pLSA and LDA. *Information Retrieval*, 14(2):178–203, 2011.

[15] T. Masada, S. Kiyasu, and S. Miyahara. Comparing lda with plsi as a dimensionality reduction method in document clustering. In *Proceedings of the 3rd LKR*, pages 13–26, 2008.

[16] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *Proceedings of NIPS*, pages 3111–3119, 2013.

[17] J. M. Ponte and W. B. Croft. A language modeling approach to information retrieval. In *Proceedings of the 21st ACM SIGIR*, pages 275–281, 1998.

[18] X. Wei and W. B. Croft. LDA-based document models for ad-hoc retrieval. In *Proceedings of the 29th ACM SIGIR*, pages 178–185, 2006.

[19] J. Xu and W. B. Croft. Query expansion using local and global document analysis. In *Proceedings of the 19th ACM SIGIR*, pages 4–11, 1996.

[20] W. Xu, X. Liu, and Y. Gong. Document clustering based on non-negative matrix factorization. In *Proceedings of the 26th ACM SIGIR*, pages 267–273, 2003.

[21] X. Yi and J. Allan. A comparative study of utilizing topic models for information retrieval. In *Proceedings of the 31st ECIR*, pages 29–41. 2009.

[22] G. Zuccon, B. Koopman, P. Bruza, and L. Azzopardi. Integrating and evaluating neural word embeddings in information retrieval. In *Proceedings of the 20th ADCS*, pages 1–8, 2015.