# Are Semantically Coherent Topic Models Useful for Ad Hoc Information Retrieval?

**Romain Deveaud** **Eric SanJuan**
University of Avignon - LIA
Avignon, France
romain.deveaud@univ-avignon.fr
eric.sanjuan@univ-avignon.fr

**Patrice Bellot**
Aix-Marseille University - LSIS
Marseille, France
patrice.bellot@lsis.org

## Abstract

The current topic modeling approaches for Information Retrieval do not allow to explicitly model query-oriented latent topics. More, the semantic coherence of the topics has never been considered in this field. We propose a model-based feedback approach that learns Latent Dirichlet Allocation topic models on the top-ranked pseudo-relevant feedback, and we measure the semantic coherence of those topics. We perform a first experimental evaluation using two major TREC test collections. Results show that retrieval performances tend to be better when using topics with higher semantic coherence.

## 1 Introduction

Representing documents as mixtures of "topics" has always been a challenge and an objective for researchers working in text-related fields. Based on the words used within a document, topic models learn topic level relations by assuming that the document covers a small set of concepts. Learning the topics from a document collection can help to extract high level semantic information, and help humans to understand the meaning of documents. Latent Semantic Indexing (Deerwester et al., 1990) (LSI), probabilistic Latent Semantic Analysis (Hofmann, 2001) (pLSA) and Latent Dirichlet Allocation (Blei et al., 2003) (LDA) are the most famous approaches that tried to tackle this problem throughout the years. Topics produced by these methods are generally fancy and appealing, and often correlate well with human concepts. This is one of the reasons of the intensive use of topic models (and especially LDA) in current research in Natural Language Processing (NLP) related areas.

One main problem in *ad hoc* Information Retrieval (IR) is the difficulty for users to translate a complex information need into a keyword query. The most popular and effective approach to overcome this problem is to improve the representation of the query by adding query-related "concepts". This approach mostly relies on pseudo-relevance feedback, where these so-called "concepts" are the most frequent words occurring in the top-ranked documents retrieved by the retrieval system (Lavrenko and Croft, 2001). From that perspective, topic models seem attractive in the sense that they can provide a descriptive and intuitive representation of concepts. But how can we quantify the usefulness of these topics with respect to an IR system? Recently, researchers developed measures which evaluate the semantic coherence of topic models (Newman et al., 2010; Mimno et al., 2011; Stevens et al., 2012). We adopt their view of semantic coherence and apply one of these measures to query-oriented topics.

Several studies concentrated on improving the quality of document ranking using topic models, especially probabilistic ones. The approach by Wei and Croft (2006) was the first to leverage LDA topics to improve the estimate of document language models and achieved good empirical results. Following this pioneering work, several studies explored the use of pLSA and LDA under different experimental settings (Park and Ramamohanarao, 2009; Yi and Allan, 2009; Andrzejewski and Buttler, 2011; Lu et al., 2011). The reported results suggest that the words and the probability distributions learned by probabilistic topic models are effective for query expansion. The main drawback of these approaches is that topics are learned on the whole target document collection prior to retrieval, thus leading to a static topical representation of the collection. Depending on the query and on its specificity, topics may either be too coarse or too fine to accurately represent the latent concepts of the query. Recently, Ye et al. (2011) proposed a method which uses

LDA and learns topics directly on a limited set of documents. While this approach is a first step towards modeling query-oriented topics, it lacks some theoretic principles and only aims to heuristically construct a "best" topic (from all learned topics) before expanding the query with its most probable words. More, none of the aforementioned works studied the semantic coherence of those generated topics. We tackle these issues by making the following contributions:

- we introduce Topic-Driven Relevance Models, a model-based feedback approach (Zhai and Lafferty, 2001) for integrating topic models into relevance models by learning topics *on* pseudo-relevant feedback documents (as opposed to the entire document collection),
- we explore the coherence of those generated topics using the queries of two major and well-established TREC test collections,
- we evaluate the effects coherent topics have on *ad hoc* IR using the same test collections.

## 2 Topic-Driven Relevance Models

### 2.1 Relevance Models

The goal of relevance models is to improve the representation of a query $Q$ by selecting terms from a set of initially retrieved documents (Lavrenko and Croft, 2001). As the concentration of relevant documents is usually higher in the top ranks of the ranking list, this is constituted by a number $N$ of top-ranked documents. Relevance models usually perform better when combined with the original query model (or maximum likelihood estimate). Let $\tilde{\theta}_Q$ be this maximum likelihood query estimate and $\hat{\theta}_Q$ a relevance model, the updated new query model is given by:

$$P(w|\theta_Q) = \lambda\, P(w|\tilde{\theta}_Q) + (1-\lambda)P(w|\hat{\theta}_Q) \quad (1)$$

where $\lambda \in [0,1]$ is a parameter that controls the tradeoff between the original query model and the relevance model. One of the most robust variants of the relevance models is computed as follows:

$$P(w|\hat{\theta}_Q) \propto \sum_{\theta_D \in \Theta} P(\theta_D) P(w|\theta_D) \prod_{t \in Q} P(t|\theta_D) \quad (2)$$

where $\Theta$ is a set of pseudo-relevant feedback documents and $\theta_D$ is the language model of document $D$. This notion of estimating a query model is often referred to as model-based feedback (Zhai and Lafferty, 2001). We assume $P(\theta_D)$ to be uniform, resulting in an estimated relevance model based on a sum of document models weighted by the query likelihood score. The final, interpolated, estimate expressed in equation (1) is often referred in the literature as RM3. We tackle the null probabilities problem by smoothing the document language model using the well-known Dirichlet smoothing (Zhai and Lafferty, 2004).

### 2.2 LDA-based Feedback Model

The estimation of the feedback model $\hat{\theta}_Q$ constitutes the first contribution of this work. We propose to explicitly model the latent topics (or concepts) that exist behind an information need, and to use them to improve the query representation. We consider $\Theta$ as the set of pseudo-relevant feedback documents from which the latent concepts would be extracted. The retrieval algorithm used to obtain these documents can be of any kind, the important point is that $\Theta$ is a reduced collection that contains the top documents ranked by an automatic and state-of-the-art retrieval process.

Instead of viewing $\Theta$ as a set of document language models that are likely to contain topical information about the query, we take a probabilistic topic modeling approach. We specifically focus on Latent Dirichlet Allocation (LDA), since it is currently one of the most representative. In LDA, each topic multinomial distribution $\phi_k$ is generated by a conjugate Dirichlet prior with parameter $\beta$, while each document multinomial distribution $\theta_d$ is generated by a conjugate Dirichlet prior with parameter $\alpha$. In other words, $\theta_{d,k}$ is the probability of topic $k$ occurring in document $D$ (i.e. $P(k|D)$). Respectively, $\phi_{k,w}$ is the probability of word $w$ belonging to topic $k$ (i.e. $P(w|k)$). We use variational inference implemented in the LDA-C software[1] to overcome intractability issues (Blei et al., 2003; Griffiths and Steyvers, 2004). Under this setting, we compute the topic-driven estimation of the query model using the following equation:

$$P(w|\hat{\theta}_Q) \propto \sum_{\theta_D \in \Theta} \Big( P(\theta_D) P(w|\theta_D) \; P_{TM}(w|D) \prod_{t \in Q} P(t|\theta_D) \Big) \quad (3)$$

where $P_{TM}(w|D)$ is the probability of word $w$ occurring in document $D$ using the previously

[1] www.cs.princeton.edu/~blei/lda-c

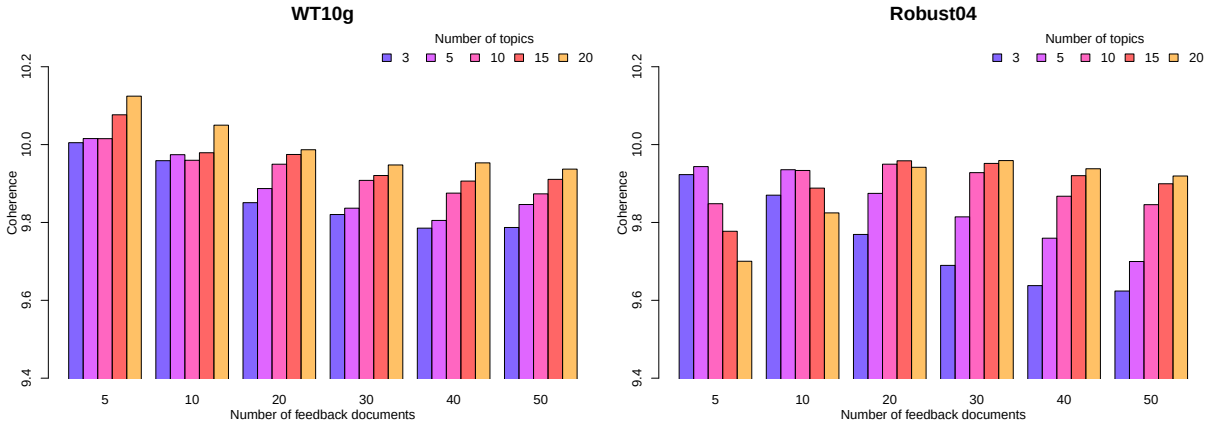


Figure 1: Semantic coherence of the topic models for different values of $K$, in function of the number $N$ of feedback documents.

learned multinomial distributions. Let $\mathcal{T}_\Theta$ be a topic model learned on the $\Theta$ set of feedback documents, this probability is given by:

$$P_{TM}(w|D) = \sum_{k \in \mathcal{T}_\Theta} \phi_{k,w} \cdot \theta_{D,k} \qquad (4)$$

High probabilities are thus given to words that are important in topic $k$, when $k$ is an important topic in document $D$. In the remainder of this paper, we refer to this general approach as TDRM for Topic-Driven Relevance Models.

### 2.3 Measuring the coherence of query-oriented topics

TDRM relies on two important parameters: the number of topics $K$ that we want to learn, and the number of feedback documents $N$ from which LDA learns the topics. Varying these two parameters can help to capture more information and to model finer topics, but how about their global semantic coherence?

Term similarities measured in restricted domains was the first step for evaluating semantic coherence (Gliozzo et al., 2007), and was a first basis for the development of several topic coherence evaluation measures (Newman et al., 2010). Computing the Pointwise Mutual Information (PMI) of all word pairs over Wikipedia was found to be an effective metric using news and books corpora. Recently, Stevens et al. (2012) used (among others) an aggregate version of this metric to evaluate large amounts of topic models. We use this method to evaluate the coherence of query-oriented topics. Specifically, the coherence of a topic model $\mathcal{T}_\Theta^K$ composed of $K$ topics is:

$$c(\mathcal{T}_\Theta^K) = \frac{1}{K} \sum_{i=1}^{K} \sum_{(w,w') \in k_i} \log \frac{P(w,w') + \epsilon}{P(w)P(w')} \qquad (5)$$

where probabilities of word occurrences and co-occurrences are estimated using an external reference corpus. Following Newman et al. (2010), we use Wikipedia to compute PMI and set $\epsilon = 1$ as in (Stevens et al., 2012).

## 3 Evaluation

### 3.1 Experimental setup

We performed our evaluation using two main TREC[2] collections: Robust04 and WT10g. Robust04 is composed 528,155 of news articles coming from three newspapers and the FBIS. It supported the TREC 2004 Robust track, from which we used the 250 query topics (numbers: 301-450, 601-700). The WT10g collection is composed of 1,692,096 web pages, and supported the TREC Web track for four years (2001-2004). We focus on the 2000 and 2001 ad-hoc query topics (numbers: 451-550). We used the open-source indexing and retrieval system Indri[3] to run our experiments. We indexed the two collections with the exact same parameters: tokens were stemmed with the well-known light Krovetz stemmer and stopwords were removed using the standard English stoplist embedded with Indri (417 words).

### 3.2 Semantic coherence evaluation

Most coherent topics are composed of rare words that do not often occur in the reference corpus, but

[2] trec.nist.gov
[3] lemurproject.org/indri.php

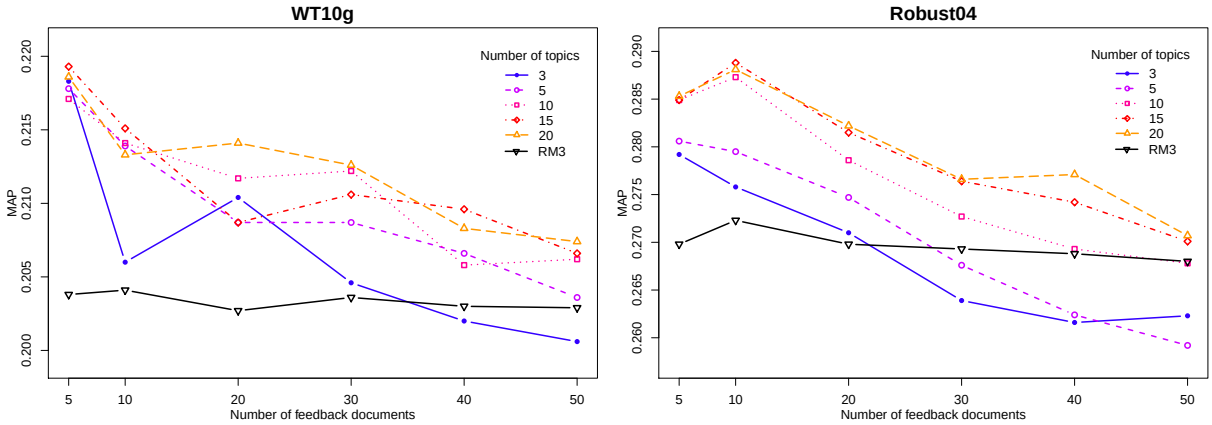


Figure 2: Retrieval performance in terms of Mean Average Precision (MAP) of the TDRM approach. Each line represent a different number of topics $K$, and the performance are reported in function the number $N$ of feedback documents. The black, plain line represents the RM3 baseline.

co-occur at lot together. We see on Figure 1 that very coherent topics are identified in the top 5 and 10 feedback documents for the WT10g collection, suggesting that closely related documents are retrieved in the top ranks. Results are quite different on the Robust04 collection, where topic models with 20 topics on 5 documents are the least coherent. However, when looking at the Robust04 documents, we see that they are on average almost twice smaller than the WT10g web pages. We hypothesize that the heterogeneous nature of the web allows to model very different topics covering several aspects of the query, while news articles are contributions focused on a single subject.

Overall, the more coherent topic models contain a reasonable amount of topics (10-15), thus allowing to fit with variable amounts of documents. The attentive reader will notice that the topic coherence scores are very high compared to those previously reported in the literature (Stevens et al., 2012). The TDRM approach captures topics that are centered around a specific information need, often with a limited vocabulary, which favors word co-occurrence. On the other hand, topics learned on entire collections are coarser than ours, which leads to lower coherence scores.

### 3.3 Document retrieval results

Since TDRM is based on Relevance Models (Lavrenko and Croft, 2001), we take the RM3 approach presented in Section 2.1 as baseline. The $\lambda$ parameter is common between RM3 and TDRM and is determined for each query using leave-one-query-out cross-validation (that is: learn the best parameter setting for all queries but one, and evaluate the held-out query using the previously learned parameter).

We report *ad hoc* document retrieval performances in Figure 2. We noticed in the previous section that the most coherent topic models were modeled using 5 feedback documents and 20 topics for the WT10g collection, and this parameter combination also achieves the best retrieval results. Overall, using 10, 15 or 20 topics allow it to achieve high and similar performance from 5 to 20 documents. We observe than using 20 topics for the Robust04 collection consistently achieves the highest results, with the topic model coherence growing as the number of feedback documents increases. Although topics coming from news articles may be limited, they benefit from the rich vocabulary of professional writers who are trained to avoid repetition. Their use of synonyms allows TDRM to model deep topics, with a comprehensive description of query aspects. Since synonyms are less likely to co-occur in encyclopedic articles like Wikipedia, we think that, in our case, the semantic coherence measure could be more accurate using other textual resources. This measure seems however to be effective when dealing with heterogeneously structured documents.

## 4 Conclusions & Future Work

Overall, modeling query-oriented topic models and estimating the feedback query model using these topics greatly improves *ad hoc* Information Retrieval, compared to state-of-the-art relevance models. While semantically coherent topic mod-

els do not seem to be effective in the context of a news articles search task, they are a good indicator of effectiveness in the context of web search. Measuring the semantic coherence of query topics could help predict query effectiveness or even choose the best query-representative topic model.

## Acknowledgments

This work was supported by the French Agency for Scientific Research (Agence Nationale de la Recherche) under CAAS project (ANR 2010 CORD 001 02).

## References

David Andrzejewski and David Buttler. 2011. Latent Topic Feedback for Information Retrieval. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '11, pages 600–608.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3:993–1022.

Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman. 1990. Indexing by Latent Semantic Analysis. *Journal of the American Society for Information Science*, 41(6):391–407.

Alfio Massimiliano Gliozzo, Marco Pennacchiotti, and Patrick Pantel. 2007. The Domain Restriction Hypothesis: Relating Term Similarity and Semantic Consistency. In *Human Language Technologies: The 2007 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 131–138.

Thomas L Griffiths and Mark Steyvers. 2004. Finding scientific topics. *Proceedings of the National Academy of Sciences of the United States of America*, 101 Suppl.

Thomas Hofmann. 2001. Unsupervised Learning by Probabilistic Latent Semantic Analysis. *Machine Learning*, 42:177–196.

Victor Lavrenko and W. Bruce Croft. 2001. Relevance-Based Language Models. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '01, pages 120–127.

Yue Lu, Qiaozhu Mei, and ChengXiang Zhai. 2011. Investigating task performance of probabilistic topic models: an empirical study of PLSA and LDA. *Information Retrieval*, 14:178–203.

David Mimno, Hanna M. Wallach, Edmund Talley, Miriam Leenders, and Andrew McCallum. 2011. Optimizing Semantic Coherence in Topic Models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 262–272.

David Newman, Jey Han Lau, Karl Grieser, and Timothy Baldwin. 2010. Automatic Evaluation of Topic Coherence. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 100–108.

Laurence A. Park and Kotagiri Ramamohanarao. 2009. The Sensitivity of Latent Dirichlet Allocation for Information Retrieval. In *Proceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases*, ECML PKDD '09, pages 176–188.

Keith Stevens, Philip Kegelmeyer, David Andrzejewski, and David Buttler. 2012. Exploring Topic Coherence over Many Models and Many Topics. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, EMNLP-CoNLL '12, pages 952–961.

Xing Wei and W. Bruce Croft. 2006. LDA-based Document Models for Ad-hoc Retrieval. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '06, pages 178–185.

Zheng Ye, Jimmy Xiangji Huang, and Hongfei Lin. 2011. Finding a Good Query-Related Topic for Boosting Pseudo-Relevance Feedback. *JASIST*, 62(4):748–760.

Xing Yi and James Allan. 2009. A Comparative Study of Utilizing Topic Models for Information Retrieval. In *Proceedings of the 31th European Conference on IR Research on Advances in Information Retrieval*, ECIR '09, pages 29–41. Springer-Verlag.

Chengxiang Zhai and John Lafferty. 2001. Model-based Feedback in the Language Modeling Approach to Information Retrieval. In *Proceedings of the Tenth International Conference on Information and Knowledge Management*, CIKM '01, pages 403–410.

Chengxiang Zhai and John Lafferty. 2004. A Study of Smoothing Methods for Language Models Applied to Information Retrieval. *ACM Transactions on Information Systems*, 22(2):179–214.