Accurate and effective latent concept modeling for ad hoc information retrieval

Romain Deveaud [1], **Eric SanJuan** [2], **Patrice Bellot** [3]

1. *University of Glasgow, UK*
   romain.deveaud@glasgow.ac.uk
2. *Université d'Avignon et des Pays de Vaucluse, 84000, Avignon, France*
   eric.sanjuan@univ-avignon.fr
3. *Aix-Marseille Université, CNRS, LSIS UMR 7296, 13397, Marseille, France*
   patrice.bellot@univ-amu.fr

ABSTRACT. *A keyword query is the representation of the information need of a user, and is the result of a complex cognitive process which often results in under-specification. We propose an unsupervised method namely Latent Concept Modeling (LCM) for mining and modeling latent search concepts in order to recreate the conceptual view of the original information need. We use Latent Dirichlet Allocation (LDA) to exhibit highly-specific query-related topics from pseudo-relevant feedback documents. We define these topics as the latent concepts of the user query. We perform a thorough evaluation of our approach over two large ad-hoc TREC collections. Our findings reveal that the proposed method accurately models latent concepts, while being very effective in a query expansion retrieval setting.*

RÉSUMÉ. *Une requête est la représentation du besoin d'information d'un utilisateur, et est le résultat d'un processus cognitif complexe qui mène souvent à un mauvais choix de mots-clés. Nous proposons une méthode non supervisée pour la modélisation de concepts implicites d'une requête, dans le but de recréer la représentation conceptuelle du besoin d'information initial. Nous utilisons l'allocation de Dirichlet latente (LDA) pour détecter les concepts implicites de la requête en utilisant des documents pseudo-pertinents. Nous évaluons cette méthode en profondeur en utilisant deux collections de test de TREC. Nous trouvons notamment que notre approche permet de modéliser précisément les concepts implicites de la requête, tout en obtenant de bonnes performances dans le cadre d'une recherche de documents.*

KEYWORDS: *information retrieval, topic modeling, pseudo-relevance feedback, LDA, TREC.*

MOTS-CLÉS : *recherche d'information, modélisation thématique, retour de pertinence simulé, LDA, TREC.*

DOI:10.3166/DN.17.1.61-84

1. Introduction

Information retrieval is about satisfying a user's information need, usually by retrieving documents or passages from a target collection. Traditionally the user represents her information need by a query composed of a few words, or keywords, which is submitted to the retrieval system. The system considers this representation as input and attempts to match documents against the query words, thus forming an ordered list of documents ranked by their estimated relevance to the query. However, representing a complete information need with keywords may introduce ambiguity, or the user could lack the vocabulary or the core concepts needed to effectively formulate the query. More, Ingwersen stated in (Ingwersen, 1994) that "*the user's own request formulation is a representation of [her] current cognitive state concerned with an information need*". A query may not contain sufficient information if the user is searching for some topic in which he-she is not confident at all. Hence, without some kind of context, the retrieval system could simply miss some nuances or details that the user did not – or could not – provide in query. This context can take the form of interest modeling based on historic (or social) behavior, or can be composed of evidences extracted from documents (Finkelstein *et al.*, 2002; White *et al.*, 2009). The latter is better known under the "concept-based retrieval" idiom and received much attention throughout the years (Bai *et al.*, 2007; Bendersky *et al.*, 2011; Chang *et al.*, 2006; Egozi *et al.*, 2011; Metzler, Croft, 2007). The basic idea is to expand the queries with sets of words or multiword terms extracted from feedback documents. This feedback set is composed of documents that are relevant or pseudo-relevant to the initial query and are likely to carry important pieces of information about the search context. Words that convey the most information or that are the most relevant to the initial query are considered as latent concepts (or implicit query concepts), and used to reformulate the query.

The problem with this concept-based retrieval approach is that each word accounts for a specific concept. However, different words associations can lead to different concepts, or express different notions, that would not exist when considering words separately. Moreover, a concept represents a notion and can be viewed as a coherent fragment of knowledge. Stock (2010) gives a definition that follows that direction by asserting that a "concept *is defined as a class containing certain objects as elements, where the objects have certain properties"*. Faceted Topic Retrieval (Carterette, Chandar, 2009) is an attempt to retrieve documents that cover all the concepts (or facets) of the query. However, while assuming that a query can be related to a finite number of facets, the authors did not address the problem of query facets identification, which we tackle in this work.

The goal of this work is to accurately represent the underlying core concepts involved in a search process, hence indirectly improving the contextual information. For this purpose, we introduce an unsupervised framework that tracks the implicit concepts related to a given query, and improves query representation by incorporating these concepts to the initial query. For each query, our method extracts latent concepts from a reduced set of feedback documents initially retrieved by the system. These

documents can come from any textual source of information. The view of a concept introduced by Stock (2010) is coherent with the topics identified by topic modeling algorithms. 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.

The example presented in Table 1 shows the latent concepts identified by our approach for the query "*dinosaurs*", using a large web crawl as source of information. Each concept k is composed of words w that are topically related and weighted by their probability $P(w|k)$ of belonging to that concept. This weighting scheme emphasizes important words and effectively reflects their influence within the concept. We perform the concept extraction part using the LDA generative probabilistic model. Given a document collection, LDA computes the topic distributions over documents and the word distributions over topics. Here, we use this latter distribution to represent search-related concepts. In other words, we assimilate the topics identified by LDA in the top-ranked pseudo-relevant documents as the latent concepts of the query. Our method also weights concepts to reflect their importance with regard to the query, as further detailed in Section 3.4. Concepts that contain words that are less likely to occur in the collection will be assigned a lower weight. In our example, the words that compose the "*toys*" concept co-occur at a lower frequency than the other concepts. The weight $\hat{\delta}_2$ ($= 0.021$) reflects the rather low likelihood that the concept k_2 would be representing the underlying information need. Despite this low weight, the system would however be able to retrieve relevant documents in case the user was really searching for dinosaur toys.

The main strength of our approach is that it is entirely unsupervised and does not require any training step. The number of needed feedback documents as well as the optimal number of concepts are automatically estimated at query time. We emphasize that the algorithms have no prior information about these concepts. The method is also entirely independent of the source of information used for concept modeling. Queries are not labeled with topics or keywords and we do not manually fix any parameter at any time, except the number of words composing the concepts. We thoroughly evaluate our approach on two main TREC collections. The experimental results show that using such concepts to reformulate the query can lead to significant improvements in the document retrieval effectiveness.

The remainder of this paper is organized as follows. In Section 2, we review related topic modeling approaches for information retrieval. Section 3 provides a

Table 1. *Concepts identified for the query "dinosaurs" (TREC Web Track topic 14) by our approach. Probabilities act as weights and reflect the relative informativeness of words within a concept k. Concepts are also weighted accordingly. We have labeled the concepts manually for clarity purpose*

k_0		k_1		k_2		k_3	
$P(w\|k_0)$	word w	$P(w\|k_1)$	word w	$P(w\|k_2)$	word w	$P(w\|k_3)$	word w
0.196	feathers	0.257	dinosaur	0.370	dinosaur	0.175	dinosaur
0.130	birds	0.180	devil	0.165	price	0.125	kenya
0.112	evolved	0.095	moon-boy	0.112	party	0.122	years
0.102	flight	0.054	bakker	0.053	birthday	0.087	fossils
0.093	dinosaurs	0.054	world	0.039	game	0.082	paleontology
0.084	protopteryx	0.049	series	0.023	toys	0.072	expedition
0.065	fossil	0.045	marvel	0.021	t-rex	0.070	discovery
	...		...		...		...
birds ($\hat{\delta}_0 = 0.434$)		**comic** ($\hat{\delta}_1 = 0.254$)		**toys** ($\hat{\delta}_2 = 0.021$)		**paleontology** ($\hat{\delta}_3 = 0.291$)	

quick overview of Latent Dirichlet Allocation, then details our proposed approach. Section 4.4 gives some insights on the general sources of information we use to model latent concepts. We evaluate our approach and discuss the results in Section 4. Finally, Section 5 concludes the paper and offers some perspectives for future work.

2. Related work

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, Croft, 2001). From that perspective, topic models seem attractive in the sense that they can provide a descriptive and intuitive representation of concepts.

The work presented in this paper crosses the bridge between extra-corpora implicit feedback approaches and cluster-based information retrieval. Probabilistic topic modeling (and especially Latent Dirichlet Allocation) for information retrieval has been widely used recently in several ways (Andrzejewski, Buttler, 2011; Lu *et al.*, 2011; Park, Ramamohanarao, 2009; Wei, Croft, 2006; Yi, Allan, 2009) and all studies reported improvements in document retrieval effectiveness. The main idea is to build a static topic model (using either LSA, pLSA, or LDA) of the collection, which will never be further updated, and to smooth the document language model by incorporating probabilities of words that belong to some topics matching the query (Lu *et al.*, 2011; Park, Ramamohanarao, 2009; Wei, Croft, 2006; Yi, Allan, 2009). The idea of using feedback documents was explored in (Andrzejewski, Buttler, 2011), where query-specific topics are chosen from the top two documents returned by the original query. These topics are identified using the document-topic mixture weights previ-

ously computed by LDA over the entire collection with the aim of finally expanding the query. The main drawback of all the aforementioned 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. In contrast, our approach use topic models directly on pseudo-relevant feedback documents, which are topically related to the query. To our knowledge, our approach is the first attempt to apply probabilistic topic models to a limited set of documents in order to exhibit latent search concepts. This is also the first one to report of using several sources of feedback documents for varying the concept representations.

The impact of using different external sources of knowledge for improving retrieval effectiveness has also been investigated in the past, but studies mainly concentrated on demonstrating how the use of a single resource could improve performance. Data sources like Wikipedia (Li *et al.*, 2007; Suchanek *et al.*, 2007), WordNet (Liu *et al.*, 2004; Suchanek *et al.*, 2007), news corpora or even the Web itself (Diaz, Metzler, 2006) were used separately for enhancing search performances. Diaz and Metzler investigated in (Diaz, Metzler, 2006) the use of large and general external resources. They present a Mixture of Relevance Models that estimates the query model using a news corpus and two web corpora as external sources.

In this paper, we extend our previous studies around the Latent Concept Modeling framework (Deveaud *et al.*, 2013b), which mainly consists at applying topic modeling algorithms such as LDA to a small set of pseudo-relevant feedback documents (Deveaud *et al.*, 2013a; Ye *et al.*, 2011). While we recall the main principles of our method in the next section, we perform a thorough evaluation of the estimated parameters and of the retrieval effectiveness.

3. Latent Concept Modeling (LCM)

We propose to model the latent concepts that exist behind an information need and to use them to improve the query representation, thus leading to better retrieval. Let $\mathcal{R}$ be a collection of text documents in which the latent concepts will be extracted. An initial subset $\mathcal{R}_Q$ is formed by the top feedback documents retrieved by a first retrieval step using the initial query Q. The retrieval algorithm can be of any kind, the important point is that $\mathcal{R}_Q$ is a reduced collection that contains the top documents ranked by an automatic and state-of-the-art retrieval process.

3.1. Latent Dirichlet Allocation

Latent Dirichlet Allocation is a generative probabilistic topic model (Blei *et al.*, 2003). The underlying intuition is that documents exhibit multiple *topics*, where a *topic* is a multinomial distribution over a fixed vocabulary W: LDA considers documents as mixtures of topics and topics as mixtures of words. The goal of LDA is to

automatically discover the topics from a collection of documents. The documents of the collection are modeled as mixtures over K topics, each of which is a multinomial distribution over W. Each topic multinomial distribution ϕ_k is generated by a conjugate Dirichlet prior with parameter $\vec{\beta}$, while each document multinomial distribution θ_d is generated by a conjugate Dirichlet prior with parameter $\vec{\alpha}$. Thus, the topic proportions for document d are θ_d, and the word distributions for topic k are ϕ_k. In other words, $\theta_{d,k}$ is the probability of topic k occurring in document d (i.e. $P_{TM}(k|d)$). Respectively, $\phi_{k,w}$ is the probability of word w belonging to topic k (i.e. $P_{TM}(w|k)$). Exact LDA estimation was found to be intractable and several approximations have been developed (Blei *et al.*, 2003; Griffiths, Steyvers, 2004). We use in this work the algorithm implemented and distributed by Pr. Blei[1].

The advantage of using LDA on a query-based set of documents is that it can model topics that are highly related to the query: namely the latent concepts of the query. There are several issues that we need to tackle in order to accurately model these concepts for further retrieval. First, how to estimate the right amount of concepts? LDA is an unsupervised approach but needs some parameters, including the number of desired topics. A dozen feedback documents clearly cannot address hundreds of topics, we thus need to estimate the right amount of topics. Similarly, which number of feedback documents must be chosen to ensure that the concepts we extract are actually related to the query? In other words: how to ultimately avoid noisy concepts? Third, the different concepts do not have the same influence with respect to a given information need. The same problem occurs within the concepts where some words are more important than others. Scoring and weighting these words and concepts is then essential to reflect their contextual importance. Finally, how to use these latent concepts to actually improve document retrieval? How do they cope with existing retrieval algorithm?

We describe our approach in the following subsections, where we tackle all the issues mentioned above, while a detailed evaluation is provided in Section 4.

3.2. Estimating the number of concepts

There can be a numerous amount of concepts underlying an information need. Latent Dirichlet Allocation allows to model the topic distribution of a given collection, but the number of topics is a fixed parameter. However we cannot know in advance the number of concepts that are related to a given query. We propose a method that automatically estimates the number of latent concepts based on their word distributions.

Considering LDA's topics are constituted of the n words with highest probabilities, we define an $\mathrm{argmax}[n]$ operator which produces the top-n arguments that obtain the

1. http://www.cs.princeton.edu/~blei/lda-c

n largest values for a given function. Using this operator, we obtain the set $\mathbb{W}_k$ of the n words that have the highest probabilities $P_{TM}(w|k) = \phi_{k,w}$ in topic k:

$$\mathbb{W}_k = \underset{w}{\operatorname{argmax}}[n]\ \phi_{k,w}$$

Latent Dirichlet Allocation must be given a number of topics in order to estimate topic and word distributions. Several approaches tried to tackle the problem of automatically finding the right number of LDA's topics contained in a set of documents (Arun *et al.*, 2010; Cao *et al.*, 2009). Even though they differ at some point, they follow the same idea of computing similarities (or distances) between pairs of topics over several instances of the model, while varying the number of topics. Iterations are done by varying the number of topics of the LDA model, then estimating again the Dirichlet distributions. The optimal amount of topics of a given collection is reached when the overall dissimilarity between topics achieves its maximum value.

We propose a simple heuristic that estimates the number of latent concepts of a user query by maximizing the information divergence D between all pairs (k_i, k_j) of LDA's topics. The number of concepts $\hat{K}$ estimated by our method is given by the following formula:

$$\hat{K} = \underset{K}{\operatorname{argmax}} \frac{1}{K(K-1)} \sum_{(k,k') \in \mathbb{T}_K} D(k||k') \tag{1}$$

where K is the number of topics given as a parameter to LDA, and $\mathbb{T}_K$ is the set of K topics modeled by LDA. In other words, $\hat{K}$ is the number of topics for which LDA modeled the most scattered topics. The Kullback-Leibler divergence measures the information divergence between two probability distributions. It is used in particular by LDA in order to minimize topic variation between two expectation-maximization iterations (Blei *et al.*, 2003). It has been widely used in a variety of fields to measure similarities (or dissimilarities) between word distributions (AlSumait *et al.*, 2008). Considering it is a non-symmetric measure, we use the Jensen-Shannon divergence, which is a symmetrised version of the KL divergence, to avoid obvious problems when computing divergences between all pairs of topics. It is formally written as:

$$\begin{aligned} D(k||k') = & \frac{1}{2} \sum_{w \in \mathbb{W}_k \cap \mathbb{W}_{k'}} P_{TM}(w|k) \log \frac{P_{TM}(w|k)}{P_{TM}(w|k')} + \\ & \frac{1}{2} \sum_{w \in \mathbb{W}_k \cap \mathbb{W}_{k'}} P_{TM}(w|k') \log \frac{P_{TM}(w|k')}{P_{TM}(w|k)} \end{aligned} \tag{2}$$

The word probabilities for given topics are obtained from the multinomial distributions ϕ_k. The final outcome is an estimated number of topics $\hat{K}$ and its associated topic model. The resulting $\mathbb{T}_{\hat{K}}$ set of topics is considered as the set of latent concepts modeled from a set of feedback documents. We will further refer to the $\mathbb{T}_{\hat{K}}$ set as a *concept model*.

3.3. Maximizing conceptual coherence

A problem with pseudo-relevance feedback based approaches is that non-relevant documents can be included in the set of feedback documents. This problem is much more important with our approach since it could result with learned concepts that are not related to the initial query. We mainly tackle this difficulty by reducing the amount of feedback documents. Relevant documents concentration is higher in the top ranks of the list. Thus, one simple way to reduce the probability of catching noisy feedback documents is to reduce their overall amount. However an arbitrary number cannot be fixed for all queries. Some information needs can be satisfied by only 2 or 3 documents, while others may require 15 or 20. Thus the choice of the feedback documents amount has to be automatic for each query.

Extensive work has been done on estimating optimal samples of feedback documents for query expansion (He, Ounis, 2009; Tao, Zhai, 2006). Previous research by He and Ounis (He, Ounis, 2009) however showed that there are no or very little statistical differences between doing PRF with the top pseudo-relevant feedback documents and doing RF, depending on the size of the sample. We take a different approach here and choose the less noisy concept model instead of choosing only the most relevant feedback documents. To avoid noise, we favour the concept model that is the most similar to all the other concept models computed on different samples of feedback documents. The underlying assumption is that all the feedback documents are essentially dealing with the same topics, no matter if they are 5 or 20. Concepts that are likely to appear in different models learned from various amounts of feedback documents are certainly related to query, while noisy concepts are not. We estimate the similarity between two concept models, $\mathbb{T}_{\hat{K},m}$ and $\mathbb{T}_{\hat{K},n}$, by computing the similarities between all pairs of concepts of the two models. Considering that two concept models are generated based on different number of documents, they do not share the same probabilistic space. Since their probability distribution are not comparable, computing their overall similarity can be done solely by taking concept words into account. We treat the different concepts as bags of words and use a document frequency-based similarity measure:

$$sim(\mathbb{T}_{\hat{K},m}, \mathbb{T}_{\hat{K},n}) = \sum_{k \in \mathbb{T}_{\hat{K},m}} \sum_{k' \in \mathbb{T}_{\hat{K},n}} \frac{|k \cap k'|}{|k|} \sum_{w} \log \frac{N}{df_w} \tag{3}$$

where $|k_i \cap k_j|$ is the number of words the two concepts have in common, df_w is the document frequency of w and N is the number of documents in the target collection. The initial purpose of this measure was to track novelty (i.e. minimize similarity) between two sentences (Metzler *et al.*, 2005), which is precisely our goal, except that we want to track redundancy (i.e. maximize similarity).

The final sum of similarities between each concept pairs produces an overall similarity score of the current concept model compared to all other models. Finally, the concept model that maximizes this overall similarity is considered as the best candi-

date for representing the implicit concepts of the query. In other words, we consider the top M feedback documents for modeling the concepts, where:

$$M = \underset{1 \leq m \leq 20}{\operatorname{argmax}} \sum_{1 \leq n \leq 20, n \neq m} sim(\mathbb{T}_{\hat{K},m}, \mathbb{T}_{\hat{K},n}) \tag{4}$$

In other words, for each query, the concept model that is the most similar to all other concept models is considered as the final set of latent concepts related to the user query. The results concept model $\mathbb{T}_{\hat{K},M}$ represents the latent concepts of the query, as defined by our method.

This method requires to run several LDA model and one could question the computational cost and practical feasibility. However all models are learned on a very small number of documents (typically ranging from 1 to 20), and are then a lot faster to compute than models that operate on complete collections composed of hundreds of thousands of documents.

3.4. Concept weighting

User queries can be associated with a number of underlying concepts but these concepts do not necessarily have the same importance. Since our approach only *estimates* the best model, it still could yield noisy concepts, and some concepts may also be barely relevant. Hence it is essential to emphasize appropriate concepts and to depreciate inappropriate ones. One effective way is to rank these concepts and to weight them accordingly: important concepts will be weighted higher to reflect their importance. We define the score of a concept k as:

$$\delta_k = \sum_{D \in \mathcal{R}_Q} P(Q|D) P_{TM}(k|D) \tag{5}$$

where Q is the initial query. The underlying intuition is that relevant concepts occur in top-ranked documents and have high probabilities in these documents. The probability $P_{TM}(k|D)$ of a concept k appearing in document D is given by the multinomial distribution θ previously learned by LDA.

Each concept is weighted with respect to its likelihood of representing the query, but the actual representation of the concept is still a bag of words. Concept words are the core components of the concepts and intrinsically do not have the same importance. The easier way of weighting them is to use their probability of belonging to a concept k which are learned by Latent Dirichlet Allocation and given by the multinomial distribution ϕ_k. Probabilities are normalized across all words, the weight of word w in concept k is thus computed as follows:

$$\hat{\phi}_{k,w} = \frac{\phi_{k,w}}{\sum_{w' \in \mathbb{W}_k} \phi_{k,w'}} \tag{6}$$

$$\hat{\phi}_{k,w} = \frac{P_{TM}(w|k)}{\sum_{w' \in \mathbb{W}_k} P_{TM}(w'|k)}$$

Finally, a concept learned by our latent concept modeling approach is a set of weighted words representing a facet of the information need underlying a user query. Concepts are also weighted to reflect their relative importance.

3.5. Document ranking

The previous subsections were all about modeling consistent concepts from reliable documents and modeling their relative influence. Here we detail how these concepts can be integrated in a retrieval model in order to improve ad-hoc document ranking. There are several ways of taking conceptual aspects into account when ranking documents. Here, the final score of a document D with respect to a given user query Q is determined by the linear combination of query word matches (standard retrieval) and latent concepts matches. It is formally written as follows:

$$s(Q,D) = \lambda \cdot P(Q|D) + (1-\lambda) \cdot \prod_{k \in \mathbb{T}_{\hat{K},M}} \left(\prod_{w \in \mathbb{W}_k} P(w|D)^{\hat{\phi}_{k,w}} \right)^{\hat{\delta}_k} \tag{7}$$

where $\mathbb{T}_{\hat{K},M}$ is the *concept model* that holds the latent concepts of query Q (see Section 3.4) and $\hat{\delta}_k$ is the normalized weight of concept k:

$$\hat{\delta}_k = \frac{\delta_k}{\sum_{k' \in \mathbb{T}_{\hat{K}}} \delta_{k'}} \tag{8}$$

The equation (7) is equivalent to a developed form of Relevance Models (Lavrenko, Croft, 2001), except that we reformulate the query using the weighted words of the concepts we previously identified. The $P(Q|D)$ (respectively, $P(w|D)$) probability is the likelihood of document D being observed given the initial query Q (respectively, word w). In this work we use a language modeling approach to retrieval (Lavrenko, Croft, 2001). $P(w|D)$ is thus the maximum likelihood estimate of word w in document D, computed using the language model of document D in the target collection $\mathcal{C}$. Likewise, $P(Q|D)$ is the basic language modeling retrieval model, also known as query likelihood, and can be formally written as $P(Q|D) = \prod_{w \in Q} P(w|D)$. We tackle the null probabilities problem with the standard Dirichlet smoothing since it is more convenient for keyword queries (as opposed to verbose queries) (Zhai, Lafferty, 2004), which is the case here. We fix the Dirichlet prior parameter to 1500 and do not change it at any time during our experiments. However it is important to note that this model is generic, and that the word matching function could be entirely substituted by other state-of-the-art matching function (like BM25 (Robertson, Walker, 1994) or information-based models (Clinchant, Gaussier, 2010)) without changing the effects of our latent concept modeling approach on document ranking.

4. Experiments

4.1. Experimental setup

We evaluated Latent Concept Modeling using two large TREC[2] collections: Robust04 and ClueWeb09.

Robust04 collection is composed of news articles coming from various newspapers and was used in the TREC 2004 Robust track. It is composed of well-known corpora: FT (Financial Times), FR (Federal Register 94), LA (Los Angeles Times) and FBIS (i.e. TREC disks 4 and 5, minus the Congressional Record). The test set contains 250 query topics and complete relevance judgements.

The ClueWeb09 is the largest web test collection made available to the information retrieval community at the time of this study. This collection was involved in many TREC tracks such as the Web, Blog and Million Query tracks. We consider here the category B of the ClueWeb09 (ClueWeb09-B) which is composed of approximately 50 million Web pages. For the purpose of evaluation, we use the entire set of query topics and relevance judgements of the TREC Web track.

Table 2. *Summary of the TREC test collections used for evaluation*

Name	# documents	# unique words	Mean doc. length	Topics used
Robust04	528,155	675,613	480	301-450, 601-700
ClueWeb09-B	50,220,423	87,330,765	805	1-150

We use Indri[3] for indexing and retrieval. Both collections are indexed with the exact same parameters: tokens are stemmed with the light Krovetz stemmer, and stopwords are removed using the standard English stoplist embedded with Indri. As seen in Section 3, concepts are composed of a fixed amount of weighted words. In this work, we fix the number of words belonging to a given concept to $n = 10$. Indeed, representing an LDA topic by its top-10 most probable words is a common practice and "*usually provide[s] sufficient detail to convey the subject of a topic, and distinguish one topic from another*" (Newman *et al.*, 2010).

4.2. Analysis of the estimated parameters

The methods we present allows to generate different concept models from variable sets of pseudo-relevant feedback documents. We make the assumption that increasing the number of those documents will increase the number of latent concepts. Hence, if our Latent Concept Modeling approach is effective, the amount of estimated concepts should increase as we increase the number of documents. We present in this section

2. http://trec.nist.gov
3. http://www.lemurproject.org

a first analysis of the estimated number of concepts in function of the number of pseudo-relevant feedback documents. More specifically, we estimate the value $\hat{K}$ for each query of the two test collections and for each set of feedback documents, we vary the number of feedback documents between 1 and 20. Then, we count the number of queries for which $\hat{K}$ is the same. The feedback documents come from the target collection and are obtained by querying it using the standard language modeling approach to retrieval.

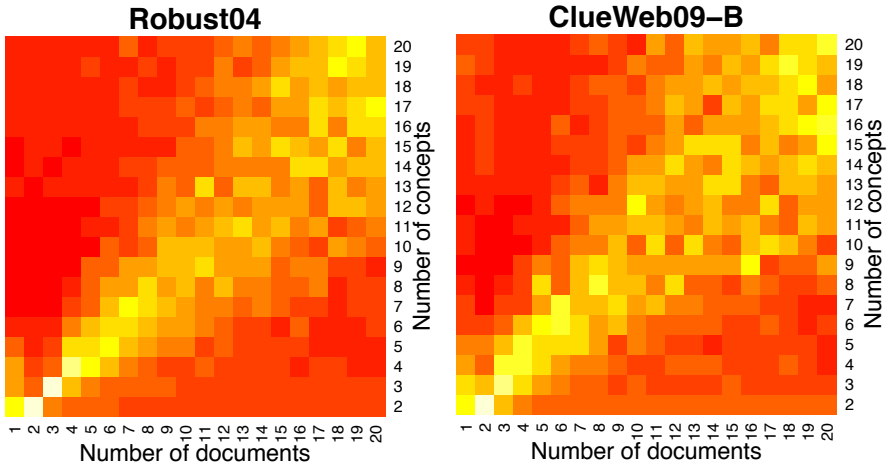

Figure 1. *Number of queries for which $\hat{K}$ where modeled, function of the number of pseudo-relevant feedback documents. A bright yellow square means a high number of queries, whereas a red square means that no query is associated to the corresponding values*

The results presented in Figure 1 show a clear correlation between the number of concepts and the number of documents. For the two collections, the great majority of queries exhibit a low amount of concepts (between 2 and 5), which are modeled from reduced set of documents (between 2 and 8). We already observe a dispersion of the number of concepts when the number of documents increases. This dispersion effect is opposed to the cohesion which can be observed for low amounts of documents. Document retrieval systems are designed to maximize (amongst others) precision at low ranks, and it is well-known that the first retrieved documents tend to be more relevant than the others. This is exactly what we see for less than 10 documents: they deal essentially with the same information, and hence the same concepts. However, increasing the number of documents increases the odds of using documents dealing with a broad range of information or of related topics, or even non-relevant documents: the dispersion cone on Figure 1 illustrates this idea. The lower part of the cone relates to the cases where many feedback documents deal with very similar information (a high number of documents and a low number of concepts), while the higher part relates to the cases where lots of topics are discussed in a few documents (with lots of noise). This effect can also be explained by the fact that some queries directly

target precise information needs, for which only a few documents can be related. All the other retrieved feedback documents may be vague and do not necessarily contain query-centric concepts.

It is also very interesting to see that the dispersion cone is more important for the ClueWeb09-B collection than for the Robust04 one (Figure 1). This is a somewhat obvious observation, but it seems that the size of the collection plays an important role on retrieving pseudo-relevant feedback documents. For the ClueWeb09-B, we see that a higher number of concepts is modeled from a limited amount of documents than for Robust04. The latter exhibits a strong correlation between the number of concepts and the number of documents, with almost a new concept appearing for each new feedback document. This is coherent with the nature of this collection, which is composed of news articles that focus in general on specific news topics. Moreover, Robust04 documents are on average two times shorter than the web pages of the ClueWeb09-B collection, which logically reduce potential topical drifts within the documents.

While we showed in this section that our approach estimates a number of concepts which is coherent with the number of feedback documents and with the nature of the test collections, we still do not know if this number of concepts really represents a “good" amount of latent search concepts. In the next section, we propose another experiment which aims at exploring the accuracy of this estimation.

4.3. Correlations with hierarchical probabilistic topic models

At this point, we cannot say whether our method can identify the “good" latent concepts of the query. Since the approach is entirely unsupervised, it learns directly from the data and does not rely on a pre-labeled training set. Therefore for each query, we do not know the number of concepts *a priori*, and *a posteriori*, we only have an estimation for which we do not know the accuracy. In this section, we propose a experiment that validates the use our method for estimating the number of concepts.

One solution to evaluate this accuracy would be to label each query manually. To do this, assessors should identify and understand all the information related to the query before extracting the concepts. Then, we could compare the concepts our method identifies with the manually identified ones. The concepts our algorithm generates are represented as bags of words with latent semantic links, but these links reveal their semantic through the interpretation of the human brain. Without humans to interpret these relations, our concepts are nothing but probability distributions over words and documents. Such a manual evaluation would thus be very subjective et would require a considerable investment.

Instead, we decided to compare the number of concepts identified by our method with the one identified by a hierarchical probabilistic topic model. More precisely, we use the Hierarchical Dirichlet Processes (Teh *et al.*, 2006) (HDP) which generalizes the Latent Dirichlet Allocation by attributing weights to concepts. We build the hierarchical topic models with the exact same pseudo-relevant feedback documents

that were used for identifying the latent search concepts. The use of HDP was often promoted by saying that it is non-parametric, i.e. it does not require the number of concepts as an input. However this parameter is actually necessary to define the dimension of the Dirichlet distribution of words over concepts. The HDP model is thus parametric but one way to overcome this limitation is by only considering the x concepts that have the highest weight (above a fixed threshold). Hence we need to define another parameter which controls the minimum value of a relevant concept in order to get rid of setting the number of concepts. In this experiment, we note this threshold t and fix it empirically to $t = 0.05$. We thus ignore all the concepts which have a weight that HDP estimated being under 0.05.

Hence, we define an automatic method for identifying and quantifying concepts that we can compare with our own approach. For each query of the two collections, we measure the correlation between the number of concepts estimated by this method and number of concepts estimated by ours. The initialization of HDP is random, the results can then change between two executions. To overcome this problem we build 10 HDP models for each set of feedback documents and keep the mean number of concepts. We report on Figure 2 the results of the Kendall τ correlation coefficient for each individual query, for the two collections. We also report mean correlations in Table 3.

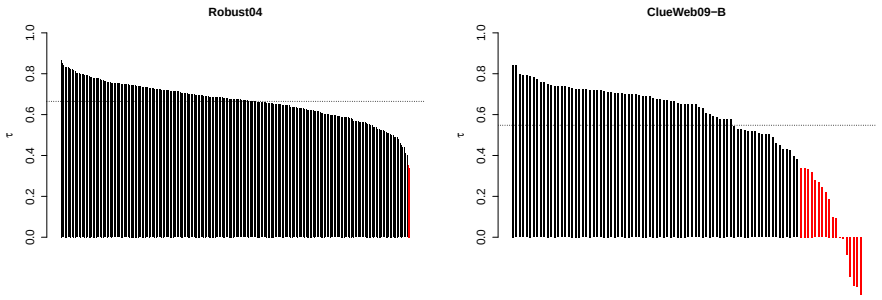

Figure 2. *Kendall's τ rank correlation coefficient between the number of topics estimated by the method presented in section 3.2 and Hierarchical Dirichlet Processes (with a threshold $t = 0.05$), for each query of each collection. Correlations represented by black bars are statistically significant at a confidence level of 95%, while red bars indicate there is no statistically significant correlation. Dotted line represents the average correlation. We ordered queries by descending correlation*

First, we can see that correlations are significant for a very large majority of queries. We notice again that modeling concepts on the ClueWeb09-B seems to be more difficult than on the Robust04: low correlations occur for 18%. The correlation coefficients for the other queries are however close to the Robust04 ones.

Overall, these results are very good and confirm that our method is capable of estimating a realistic and accurate number of concepts, correlated with the HDP model. However despite these correlations, there are great differences in the estimated num-

Table 3. *Mean correlations expressed by Pearson's ρ and Kendall's τ coefficients*

	ρ	τ
Robust04	0,782	0,665
ClueWeb09-B	0,657	0,548

bers. More precisely, the HDP method always models between 4 and 6 concepts, no matter the amount of pseudo-relevant feedback documents. Most of the time, using only one feedback document leads to a model composed of 4 concepts, while using 20 documents leads to a model with approximately 6 concepts (remind that we average over 10 models, these numbers are thus approximate). We tried to vary the threshold t but it did not have a significant impact on the results. The HDP method thus seems to identify "flat" and very sparse concepts when we use a low amount of documents. When we increase the number of documents, concepts become too general and do not capture semantic specificities. However, as we saw in the previous section, our method adapts to the quantity of information expressed in the pseudo-relevant feedback documents, and the number of concepts increases almost linearly as we add more of these information.

4.4. Identifying concepts from general sources of information

The global approach described in this paper requires a source of information from which the feedback documents could be extracted. This source of information can come from the target collection, like in traditional relevance feedback approaches, or from an external collection. In this work, we use a set of different data sources that are large enough to deal with a broad range of topics: Wikipedia as an encyclopedic source, the New York Times and GigaWord corpora as sources of news data and the category B of the ClueWeb09[4] collection as a Web source. The English GigaWord LDC corpus consists of 4,111,240 newswire articles collected from four distinct international sources including the New York Times (Graff, Cieri, 2003). The New York Times LDC corpus contains 1,855,658 news articles published between 1987 and 2007 (Sandhaus, 2008). The Wikipedia collection is a dump from July 2011 of the online encyclopedia that contains 3,214,014 documents[5]. We removed the spammed documents from the category B of the ClueWeb09 according to a standard list of spams for this collection[6]. We followed authors recommendations (Cormack *et al.*, 2011) and set the "spamminess" threshold parameter to 70. The resulting corpus is composed of 29,038,220 pages.

4. http://boston.lti.cs.cmu.edu/clueweb09/
5. http://dumps.wikimedia.org/enwiki/20110722/
6. http://plg.uwaterloo.ca/~gvcormac/clueweb09spam/

Table 4. *Information about the four general sources of information we used*

Resource	# documents	# unique words	# total words
NYT	1,855,658	1,086,233	1,378,897,246
Wiki	3,214,014	7,022,226	1,033,787,926
GW	4,111,240	1,288,389	1,397,727,483
Web	29,038,220	33,314,740	22,814,465,842

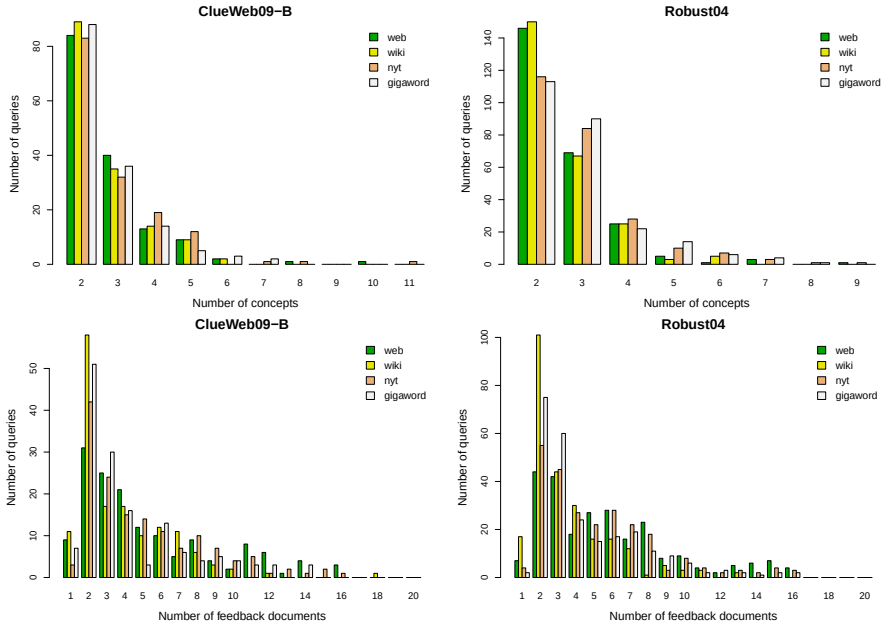

Figure 3. *Histograms that show the number of queries in function of the number $\hat{K}$ of latent concepts (Section 3.2) and the number M of feedback documents (Section 3.3)*

Figure 3 depicts the number of queries in function of the estimated numbers of latent concepts and feedback documents, for both collections. We observe that parameter estimation behaves roughly the same for the two collections. Between two and three concepts are identified for a large majority of queries. Likewise, these concepts are identified within a reduced set of between two and four documents. It is interesting to note the differences between the Web source and Wikipedia, especially for the number of feedback documents. We see that 2 or 3 Wikipedia articles are enough for approximately 60% of queries, whereas a larger number is required for the Web source. This is very coherent with the nature of Wikipedia, where articles are written with the aim of being precise and concise. When articles become too large, they are often split into several other articles that focus on very specific points. This is confirmed by a strong and statistically significant correlation between the number of

concepts $\hat{K}$ and the number of documents M for Wikipedia. Pearson's test $\rho = 0.7$ for ClueWeb09 queries, and $\rho = 0.616$ for Robust04 queries. Likewise, our method handles the heterogeneous nature of the Web and needs to choose a larger number of feedback documents in order to accurately model the latent concepts. The two parameters are less correlated for the Web source ($\rho = 0.33$ for ClueWeb09 and $\rho = 0.39$ for Robust 04), which reflects this heterogeneity and the difficulty to estimate the parameters.

4.5. LCM-based retrieval

After showing that the approach described in this paper is capable of adaptively modeling accurate concepts without relying on fixed sources of information, we evaluate in this section how they can be used to improve document retrieval effectiveness. We use three standard evaluation metrics for comparing the approaches: normalized Discounted Cumulative Gain and precision, both at 20 documents (nDGC20 and P@20), and Mean Average Precision (MAP) of the whole ranked list.

Table 5. *Document retrieval performances on two major TREC test collections. Latent concepts are modeled by the approach presented in this paper, and used to reformulate the initial user query. We use two sided paired wise t-test to determine statistically significant differences with Markov Random Field for IR (Metzler, Croft, 2005) (** : $p < 0.1$; ** : $p < 0.05$; *** : $p < 0.01$*) and Latent Concept Expansion (Metzler, Croft, 2007) (*$_{r}$: $p < 0.1$; $_{rr}$: $p < 0.05$; $_{rrr}$: $p < 0.01$*)*

	ClueWeb09-B			**Robust04**		
	nDCG@20	P@20	MAP	nDCG@20	P@20	MAP
MRF	0.2128	0.2838	0.1401	0.4231	0.3612	0.2564
LCE	0.2368	0.3095	0.1413	0.4251	0.3725*	0.2764***
GW	0.2098	0.2782	0.1283	0.4521$^{***}_{rrr}$	0.3841$^{**}_{rr}$	0.2820***
Wiki	0.2142	0.2980	0.1408	0.4189	0.3549	0.2632
NYT	0.2144	0.2816	0.1346	**0.4589**$^{***}_{rrr}$	**0.3928**$^{***}_{rrr}$	**0.2891**$^{***}_{rr}$
Web	**0.2529*****	**0.3328*****	**0.1474**	0.4428$^{*}_{r}$	0.3754*	0.2760***
Comb	0.2465***	0.3247***	0.1597$^{***}_{rrr}$	0.4680$^{***}_{rrr}$	0.3969$^{***}_{rrr}$	0.2929$^{***}_{rrr}$

Document retrieval results for both test collections are presented in Table 5. The concepts are modeled following the latent concept modeling (LCM) approach we presented. They are given a weight equal to the initial query ($\lambda = 0.5$ in equation 7). We present the results achieved when choosing each four resources separately for modeling the concepts. These approaches are compared to two competitive baselines. The first one is the Markov Random Field (MRF) model for IR, a strong baseline introduced in (Metzler, Croft, 2005) which models adjacent query terms dependencies and performs proximity search. The second one is the Latent Concept Expansion model (Metzler, Croft, 2007), which expands the initial query with the top informative single term concepts extracted from the top pseudo-relevant documents. For both baselines (MRF and LCE), we follow author's recommendation and set the weights to 0.85, 0.10 and 0.05 for query terms, bigram and proximity matches respectively.

These approaches are known for having performed consistently well amongst various test collections, including those used in our experiments (Clarke *et al.*, 2011; Metzler, Croft, 2007).

We see in Table 5 that results vary a lot depending on the resource used for concept modeling. For Web search (with the ClueWeb09 collection), the GigaWord, the New York Times and Wikipedia are not consistent at providing high quality concepts. The best results amongst these three are achieved either by the New York Times or by Wikipedia, and they perform roughly at the same level as MRF does. On the other side, the Web resource achieves higher results that are statistically significant compared to the two baselines, except for MAP. For news search (with the Robust04 collection), the influence of the four resources is clearly different. We see that the best and most statistically significant results are achieved when using concepts modeled from the NYT and the GigaWord, which are news sources, while the Web resource also performs well.

The nature of the resource from which concepts are modeled seems to be highly correlated with the document collection. We see indeed that the Web resource yields better concepts for Web search while the other resources fail. Similarly, news-based resources better help retrieval in a news search context. This may be due to word overlap between the resources and the collections, but the GigaWord and the NYT only share 18.7% of their unique words. They are very similar for modeling latent concepts of news-related search queries, but very different when looking at their vocabulary. The size of the Web resource plays a major role. Its results are consistent on the Robust04 collection and are statistically significant compared to the baselines. On the other side, using Wikipedia, which is the only resource that does not share its nature with a test collection, consistently failed to improve document retrieval for both search tasks. We thus assume that using Wikipedia for modeling latent concepts could be useful when searching for encyclopedic documents and we leave it for future work.

We also explored the combination of the latent concepts modeled from all the four sources together by averaging all *concept models* in the document scoring function:

$$s(Q,D) = \lambda \cdot P(Q|D) + (1-\lambda) \cdot \frac{1}{|\mathcal{S}|} \sum_{\sigma \in \mathcal{S}} \prod_{k \in \mathbb{T}^{\sigma}_{\hat{K}(M)}} \left(\prod_{w \in \mathbb{W}_k} P(w|D)^{\hat{\phi}_{k,w}} \right)^{\hat{\delta}_k} \quad (9)$$

where $\mathbb{T}^{\sigma}_{\hat{K},M}$ is the *concept model* built from the information source σ belonging to a set $\mathcal{S}$. This type of combination is similar to the Mixture of Relevance Models previously experimented by Diaz and Metzler (Diaz, Metzler, 2006). The results presented in Table 5 in the row **Comb** are not surprising and show support for the principles of *polyrepresentation* (Ingwersen, 1994) and *intentional redundancy* (Jones, 1990) which state that combining cognitively and structurally different representations of the information needs and documents will increase the likelihood of finding relevant documents. Even if the combination does not improve the results over the single best performing source of information, it always reaches the highest level of significance with respect to the baselines. Despite their low performance when used alone, “minor"

sources of information play an essential role to improve retrieval by modeling unique and coherent latent concepts that fit to the whole multiple concept model.

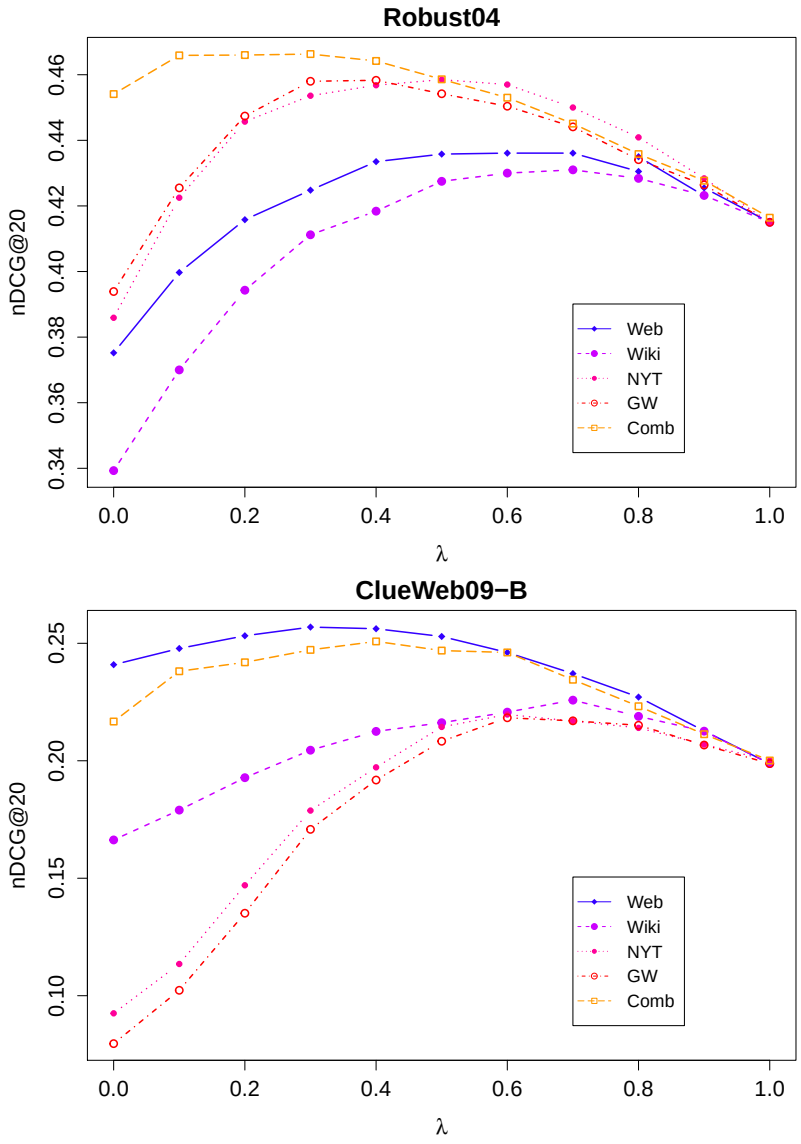

Figure 4. *Retrieval performance (nDCG@20) as a function of parameter* λ

Finally, we explore the performance of this concept-based retrieval approach by varying the λ parameter which controls the trade-off between the latent concepts and the original query. These performances are plotted in Figure 4, where high values of λ mean a high influence of the original query with respect to the latent concepts.

Unsurprisingly, best values of λ tend to be high for information sources that achieve low results, and low for information sources that achieve high results. When setting $\lambda = 0$, only the combination of information sources achieves better results than setting $\lambda = 1$ for the Robust04 collection. More, taken separately, all the concepts identified from these different sources are statistically significantly less effective than the original query. The combination of concepts modeled from heterogeneous sources is thus a better representation of the underlying information need than the original query. This results also confirm that the concepts are very different from one information source to another. However, they are not irrelevant and contribute to an accurate and complete representation of the information need.

5. Conclusion

We presented in this paper Latent Concept Modeling, a new unsupervised approach for modeling the implicit concepts lying behind a user query. These concepts are extracted from subsets of pseudo-relevant feedback documents coming from heterogeneous external resources. The number of latent concepts and the appropriate number of feedback documents are automatically estimated at query time without any previous training step. Overall, our method performs consistently well over large TREC test collections when the sources of information match the collection. The best results are achieved when combining the latent concepts modeled from all available sources of information. That shows that our approach is robust enough to handle heterogeneous documents dealing with various topics to finally model latent concepts of the query.

Our approach requires to compute many LDA models, since it jointly estimates $\hat{K}$ and M. A separate number of concepts K is estimated for each set of the top-m feedback documents, and the “best" model is chosen from the $K \times m$ matrix of models. All models, however, are learned on a very small number of documents (ranging from 1 to 20) Since the models are computed on small pieces of text (typically from 500 to 10,000 words), computation is a lot faster than for complete collections composed of millions of documents. Since long queries can take up to 5 minutes, it is clearly not feasible in the context of a live search engine. However we did not optimize the algorithms, neither did we take advantage of parallel programming. We think our approach could also benefit from future advances in the computation and estimation of LDA.

Apart from helping document retrieval, Latent Concept Modeling could be used to display intelligent, human-readable concepts in order to help the user during search. Concepts often refer to one or several entities, entity linking could then be another application of our method, as well as faceted topic retrieval. Moreover, this method is not only restricted to queries. Indeed, it can be used to quantify and model the latent concepts of any short piece of text, such as tweets or questions for example. Contextualizing tweets using sentences that cover these latent concepts was shown to be effective (Deveaud, Boudin, 2012), and is thus a promising future application.

References

AlSumait L., Barbará D., Domeniconi C. (2008). On-line LDA: Adaptive Topic Models for Mining Text Streams with Applications to Topic Detection and Tracking. In *Proceedings of the Eighth IEEE International Conference on Data Mining*, p. 3-12. IEEE Computer Society.

Andrzejewski D., Buttler D. (2011). Latent Topic Feedback for Information Retrieval. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 600–608. New York, NY, USA, ACM.

Arun R., Suresh V., Veni Madhavan C. E., Narasimha Murthy M. N. (2010). On Finding the Natural Number of Topics with Latent Dirichlet Allocation: Some Observations. In *Proceedings of the 14th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining - Volume Part I*, pp. 391–402. Berlin, Heidelberg, Springer-Verlag.

Bai J., Nie J.-Y., Cao G., Bouchard H. (2007). Using Query Contexts in Information Retrieval. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 15–22. New York, NY, USA, ACM.

Bendersky M., Metzler D., Croft W. B. (2011). Parameterized Concept Weighting in Verbose Queries. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 605–614. New York, NY, USA, ACM.

Blei D. M., Ng A. Y., Jordan M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, Vol. 3, pp. 993–1022.

Cao J., Xia T., Li J., Zhang Y., Tang S. (2009, March). A Density-based Method for Adaptive LDA Model Selection. *Neurocomputing*, Vol. 72, No. 7-9, pp. 1775–1781.

Carterette B., Chandar P. (2009). Probabilistic Models of Ranking Novel Documents for Faceted Topic Retrieval. In *Proceedings of the 18th acm conference on information and knowledge management*, pp. 1287–1296. New York, NY, USA, ACM.

Chang Y., Ounis I., Kim M. (2006). Query reformulation using automatically generated query concepts from a document space. *Information Processing & Management*, Vol. 42, No. 2.

Clarke C. L. A., Craswell N., Soboroff I., Voorhees E. M. (2011). Overview of the TREC 2011 Web Track. In E. M. Voorhees, L. P. Buckland (Eds.), *Trec*. National Institute of Standards and Technology (NIST).

Clinchant S., Gaussier E. (2010). Information-based Models for Ad Hoc IR. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, pp. 234–241. New York, NY, USA, ACM.

Cormack G. V., Smucker M. D., Clarke C. L. (2011, October). Efficient and Effective Spam Filtering and Re-ranking for Large Web Datasets. *Information Retrieval*, Vol. 14, No. 5, pp. 441–465.

Deerwester S., Dumais S. T., Furnas G. W., Landauer T. K., Harshman R. (1990). Indexing by Latent Semantic Analysis. *Journal of the American Society for Information Science*, Vol. 41, No. 6, pp. 391–407.

Deveaud R., Boudin F. (2012). LIA/LINA at the INEX 2012 Tweet Contextualization track. In *Focused Access to Content, Structure and Context: 11th International Workshop of the Initiative for the Evaluation of XML Retrieval.*

Deveaud R., SanJuan E., Bellot P. (2013a). Are Semantically Coherent Topic Models Useful for Ad Hoc Information Retrieval? In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 148–152. Sofia, Bulgaria, Association for Computational Linguistics.

Deveaud R., SanJuan E., Bellot P. (2013b). Unsupervised Latent Concept Modeling to Identify Query Facets. In *Proceedings of the 10th Conference on Open Research Areas in Information Retrieval*, pp. 93–100. Paris, France, France, Le centre de hautes études internationales d'informatique documentaire.

Diaz F., Metzler D. (2006). Improving the Estimation of Relevance Models Using Large External Corpora. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 154–161. New York, NY, USA, ACM.

Egozi O., Markovitch S., Gabrilovich E. (2011, April). Concept-Based Information Retrieval Using Explicit Semantic Analysis. *ACM Transactions on Information Systems*, Vol. 29, No. 2, pp. 8:1–8:34.

Finkelstein L., Gabrilovich E., Matias Y., Rivlin E., Solan Z., Wolfman G. (2002). Placing Search in Context: the Concept Revisited. *ACM Transactions on Information Systems*, Vol. 20, No. 1, pp. 116–131.

Graff D., Cieri C. (2003). English Gigaword. *Philadelphia: Linguistic Data Consortium*, Vol. LDC2003T05. http://www.ldc.upenn.edu/Catalog/catalogEntry.jsp?catalogId=LDC2003T05

Griffiths T. L., Steyvers M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences of the United States of America*, Vol. 101 Suppl.

He B., Ounis I. (2009). Finding Good Feedback Documents. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management*, pp. 2011–2014. New York, NY, USA, ACM.

Hofmann T. (2001). Unsupervised Learning by Probabilistic Latent Semantic Analysis. *Machine Learning*, Vol. 42, pp. 177-196. http://dx.doi.org/10.1023/A%3A1007617005950

Ingwersen P. (1994). Polyrepresentation of Information Needs and Semantic Entities: Elements of a Cognitive Theory for Information Retrieval Interaction. In *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 101–110. New York, NY, USA, Springer-Verlag New York, Inc.

Jones K. (1990). *Retrieving Information Or Answering Questions?* British Library Research and Development Department.

Lavrenko V., Croft W. B. (2001). Relevance Based Language Models. In *Proceedings of the 24Th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 120–127. New York, NY, USA, ACM.

Li Y., Luk W. P. R., Ho K. S. E., Chung F. L. K. (2007). Improving Weak Ad-hoc Queries Using Wikipedia As External Corpus. In *Proceedings of the 30th annual international acm sigir conference on research and development in information retrieval*, pp. 797–798. New York, NY, USA, ACM.

Liu S., Liu F., Yu C., Meng W. (2004). An Effective Approach to Document Retrieval via Utilizing Wordnet and Recognizing Phrases. In *Proceedings of the 27th annual international*

acm sigir conference on research and development in information retrieval, pp. 266–272. New York, NY, USA, ACM.

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

Metzler D., Bernstein Y., Croft W. B., Moffat A., Zobel J. (2005). Similarity Measures for Tracking Information Flow. In *Proceedings of the 14th ACM International Conference on Information and Knowledge Management*, pp. 517–524. New York, NY, USA, ACM.

Metzler D., Croft W. B. (2005). A Markov Random Field Model for Term Dependencies. In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 472–479. New York, NY, USA, ACM.

Metzler D., Croft W. B. (2007). Latent Concept Expansion Using Markov Random Fields. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 311–318. New York, NY, USA, ACM.

Newman D., Lau J. H., Grieser K., Baldwin T. (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*, pp. 100–108. Stroudsburg, PA, USA, Association for Computational Linguistics.

Park L. A., Ramamohanarao K. (2009). The Sensitivity of Latent Dirichlet Allocation for Information Retrieval. In *Proceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases: Part II*, pp. 176–188. Berlin, Heidelberg, Springer-Verlag.

Robertson S. E., Walker S. (1994). Some Simple Effective Approximations to the 2-Poisson Model for Probabilistic Weighted Retrieval. In *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 232–241. New York, NY, USA, Springer-Verlag New York, Inc.

Sandhaus E. (2008). The New York Times Annotated Corpus. *Philadelphia: Linguistic Data Consortium*, Vol. LDC2008T19. http://www.ldc.upenn.edu/Catalog/catalogEntry.jsp?catalogId=LDC2008T19

Suchanek F. M., Kasneci G., Weikum G. (2007). Yago: A Core of Semantic Knowledge. In *Proceedings of the 16th international conference on world wide web*, pp. 697–706. New York, NY, USA, ACM.

Tao T., Zhai C. (2006). Regularized Estimation of Mixture Models for Robust Pseudo-relevance Feedback. In *Proceedings of the 29th annual international acm sigir conference on research and development in information retrieval*, pp. 162–169. New York, NY, USA, ACM.

Teh Y. W., Jordan M. I., Beal M. J., Blei D. M. (2006). Hierarchical Dirichlet Processes. *Journal of the American Statistical Association*, Vol. 101, No. 476, pp. 1566-1581.

Wei X., Croft W. B. (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*, pp. 178–185. New York, NY, USA, ACM.

White R. W., Bailey P., Chen L. (2009). Predicting User Interests from Contextual Information. In *Proceedings of the 32Nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 363–370. New York, NY, USA, ACM.

Ye Z., Huang J. X., Lin H. (2011). Finding a Good Query-Related Topic for Boosting Pseudo-Relevance Feedback. *JASIST*, Vol. 62, No. 4, pp. 748–760.

Yi X., Allan J. (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*, pp. 29–41. Berlin, Heidelberg, Springer-Verlag.

Zhai C., Lafferty J. (2004). A Study of Smoothing Methods for Language Models Applied to Information Retrieval. *ACM Transactions on Information Systems*, Vol. 22, No. 2, pp. 179–214.