Unsupervised Latent Concept Modeling to Identify Query Facets

Romain Deveaud
University of Avignon - LIA
Avignon, France
romain.deveaud@univ-avignon.fr

Eric SanJuan
University of Avignon - LIA
Avignon, France
eric.sanjuan@univ-avignon.fr

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

ABSTRACT

Translating an information need into a keyword query can be a complex cognitive process which often results in under-specification. Retrieving documents based solely on keywords can lead the user to browse documents that do not address the specific query facets she was looking for. We introduce an unsupervised method for mining and modeling latent search concepts in order to increase the coverage of these facets. We use Latent Dirichlet Allocation (LDA), a generative probabilistic topic model, to exhibit highly-specific query-related topics from pseudo-relevant feedback documents. We define these topics as the latent concepts of the user query. The main strength of our approach is that it automatically estimates the number of latent concepts as well as the needed amount of feedback documents, without any prior training step. We evaluate our approach over two large ad-hoc TREC collections, and results show that our approach significantly improves document retrieval effectiveness and even provides a better representation of the information need than the original query.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

General Terms

Experimentation

Keywords

Information retrieval, topic modeling, relevance feedback

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 tries 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 [20] 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 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 [16, 35]. The latter is better known under the "concept-based retrieval" idiom and received much attention throughout the years [5, 6, 10, 15, 27]. 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 approach of concept-based retrieval is that each word accounts for a specific concept. There is no explicit demarcation between the different latent concepts in these studies. A concept however represents a notion and can be viewed as a knowledge ensemble. Stock [32] 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 [9] 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 surrounding this search. For this purpose, we introduce an unsupervised framework that tracks the implicit concepts related to a given query, and improves query representation by incor-

OAIR'13 , May 22-24, 2013, Lisbon, Portugal.

birds		**comic**		**toys**		**paleontology**	
$P(w\|k)$	word w	$P(w\|k)$	word w	$P(w\|k)$	word w	$P(w\|k)$	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
	...		...		...		...
$\hat{\delta}_0 = 0.434$		$\hat{\delta}_1 = 0.254$		$\hat{\delta}_2 = 0.021$		$\hat{\delta}_3 = 0.291$	

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 set concept labels manually for clarity purpose.

porating 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 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 normalized 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 Latent Dirichlet Allocation [7], a 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. Our method also weighs concepts to reflect their importance w.r.t. the query. The weight $\hat{\delta}_2$ ($= 0.021$) thus reflects the low probability of the corresponding concept being the one that is actually concerned by the initial query. 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.

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 quick overview of Latent Dirichlet Allocation, then details our proposed approach. Section 4.1 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

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 [3, 24, 29, 34, 36] and all studies reported improvements in document retrieval effectiveness. The main idea is to cluster the document collection *a priori* and smooth the document language model by incorporating probabilities of words that belong to some topics matching the query [24, 34, 36]. Other approaches also tried to directly expand the query with the words that belong to these pseudo-relevant topics [3, 29]. The idea of using feedback documents was explored in [3], 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 previously computed by LDA over the entire collection with the aim of finally expanding the query. To our knowledge, our approach is the first attempt to model topics with a probabilistic topic model from a limited set of feedback documents in order to exhibit latent search concepts.

3. LATENT CONCEPT MODELING

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 textual source of information 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.

Latent Dirichlet Allocation [7] is a probabilistic topic modeling algorithm that considers documents as mixtures of topics and topics as mixtures of words. 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 concepts. 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 this section, where we tackle all the issues mentioned above, while a detailed evaluation is provided in Section 4.

3.1 Latent Dirichlet Allocation

Latent Dirichlet Allocation is a generative probabilistic topic model [7]. The underlying intuition is that documents exhibit multiple *topics*, where a *topic* is a multinomial distribution over a fixed vocabulary W. The goal of LDA is thus 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 [7, 18]. We use in this work the algorithm implemented and distributed by Pr. Blei[1].

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 argmax[n] operator which produces the top-n arguments that obtain the 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 needs a given 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 [4, 8]. 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:

$$\hat{K} = \underset{K}{\operatorname{argmax}} \frac{1}{K(K-1)} \sum_{(k_i,k_j)\in\mathbb{T}_K} D(k_i||k_j) \qquad (1)$$

where K is the number of topics given as a parameter to LDA, and $\mathbb{T}_K$ is the set of K topics. 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 [7]. It has been widely used in a variety of fields to measure similarities (or dissimilarities) between word distributions [2]. 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. 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

An obvious problem with pseudo-relevance feedback based approaches is that not-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 [19, 22, 33]. Previous research by He and Ounis [19] however showed that there are no or very little statistical differences between doing PRF with the top pseudo-relevant feedback documents and doing PRF with only relevant documents, 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 favor 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 by computing the similarities between all pairs of concepts of the two models.

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

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:

$$sim(\mathbb{T}_{\hat{K},m}, \mathbb{T}_{\hat{K},n}) = \sum_{k \in \mathbb{T}_{\hat{K},m}} \sum_{k' \in \mathbb{T}_{\hat{K},n}} \frac{|k_i \cap k'_j|}{|k_i|} \sum_{w \in W} \log \frac{N}{df_w} \quad (2)$$

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

$$M = \operatorname*{argmax}_m \sum_n sim(\mathbb{T}_{\hat{K},m}, \mathbb{T}_{\hat{K},n}) \quad (3)$$

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.

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.

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 weigh them accordingly: important concepts will be weighted higher to reflect their importance. We define the score of a concept k as $\delta_k = \sum_{D \in \mathcal{R}_Q} P(Q|D)P(k|D)$. 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:

$$\hat{\phi}_{k,w} = \frac{\phi_{k,w}}{\sum_{w' \in \mathbb{W}_k} \phi_{k,w'}} \quad (4)$$

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 adhoc 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:

$$s(Q, D) = \lambda \cdot P(Q|D) + (1-\lambda) \cdot \prod_{k \in \mathbb{T}_{\hat{K},M}} \hat{\delta}_k \prod_{w \in \mathbb{W}_k} \hat{\phi}_{k,w} \cdot P(w|D) \quad (5)$$

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:

$$\hat{\delta}_k = \frac{\delta_k}{\sum_{k' \in \mathbb{T}_{\hat{K}}} \delta_{k'}} \quad (6)$$

The $P(Q|D)$ and $P(w|D)$ probabilities are 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 [23]. $P(w|D)$ 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) [37], 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 [30] or information-based models [12]) without changing the effects of our latent concept modeling approach on document ranking.

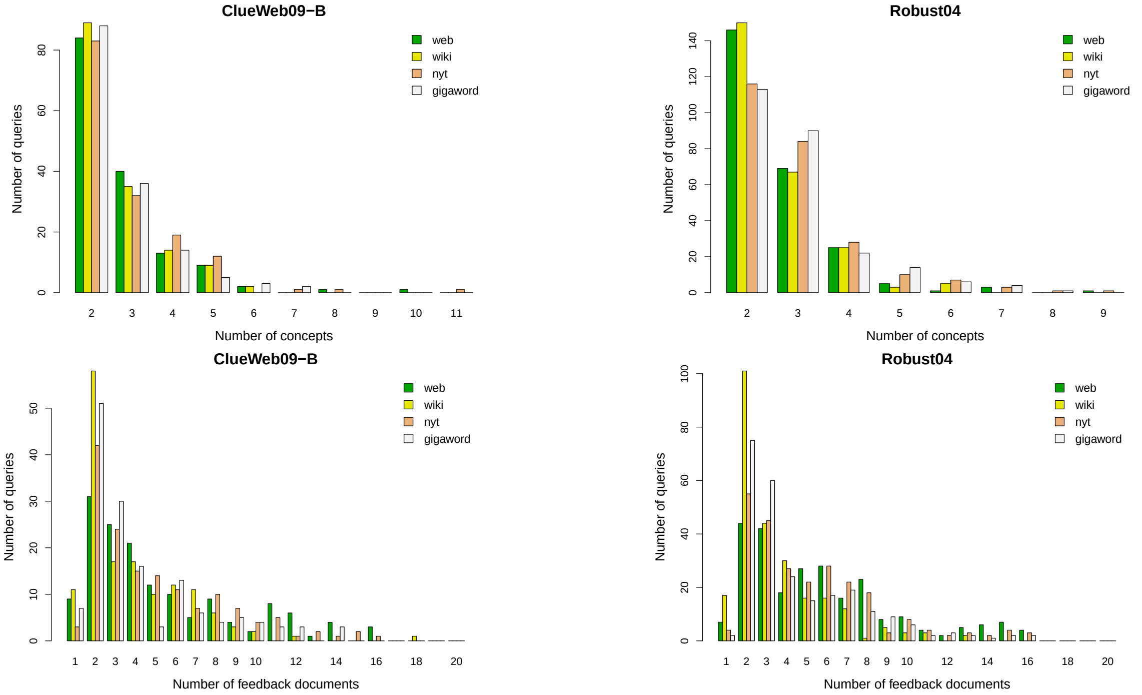

Figure 1: Histograms that show the number of queries in function of the number $\hat{K}$ of latent concepts (Section 3.2) and the number M of feedback documents (Section 3.3).

4. EVALUATION

4.1 General Sources of Information

The approach described in the previous section 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[2] 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 [17]. The New York Times LDC corpus contains 1,855,658 news articles published between 1987 and 2007 [31]. The Wikipedia collection is a dump from July 2011 of the online encyclopedia that contains 3,214,014 documents[3]. We removed the spammed documents from the category B of the ClueWeb09 according to a standard list of spams for this collection[4]. We followed authors recommendations [13] and set the "spamminess" threshold parameter to 70. The resulting corpus is composed of 29,038,220 pages.

4.2 Experimental setup

We performed our evaluation using two main TREC[5] collections. The Robust04 collection is composed of news articles coming from various newspapers and was used in the TREC 2004 Robust track. It is composed of standard 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 for the entire set. The ClueWeb09 is the largest web test collection made available to the IR 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.

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

Table 2: Information about the four general sources of information used in this work.

Name	# documents	Topics used
Robust04	528,155	301-450, 601-700
ClueWeb09-B	50,220,423	1-150

Table 4: Summary of the TREC test collections used for evaluation.

We used Indri[6] for indexing and retrieval. The collections were indexed 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. As seen in Section 3, concepts are composed of a fixed amount of weighted words. In this work,

[2] http://boston.lti.cs.cmu.edu/clueweb09/

[3] http://dumps.wikimedia.org/enwiki/20110722/

[4] http://plg.uwaterloo.ca/~gvcormac/clueweb09spam/

[5] http://trec.nist.gov

[6] http://www.lemurproject.org

	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	$\mathbf{0.4589}^{***}_{rrr}$	$\mathbf{0.3928}^{***}_{rrr}$	$\mathbf{0.2891}^{***}_{rr}$
Web	$\mathbf{0.2529}^{***}$	$\mathbf{0.3328}^{***}$	**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}

Table 3: 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 [26] ($^{*}: p < 0.1$; $^{**}: p < 0.05$; $^{***}: p < 0.01$) and Latent Concept Expansion [27] ($_{r}: p < 0.1$; $_{rr}: p < 0.05$; $_{rrr}: p < 0.01$).

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*" [28]. We use three standard evaluation metrics for comparing the approaches: nDCG and precision at 20 documents, and mean average precision (MAP) of the entire ranked list.

4.3 Analysis of the estimated parameters

Figure 1 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.

4.4 LCM-based retrieval

Document retrieval results for the two test collections are presented in Table 3. The concepts are modeled following the latent concept modeling (LCM) approach presented in this paper and are given a weight equal to the initial query ($\lambda = 0.5$ in equation 5). 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 [26] which models adjacent query terms dependencies and performs proximity search. The second one is the Latent Concept Expansion model [27], 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 [11, 27].

We see in Table 3 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. 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

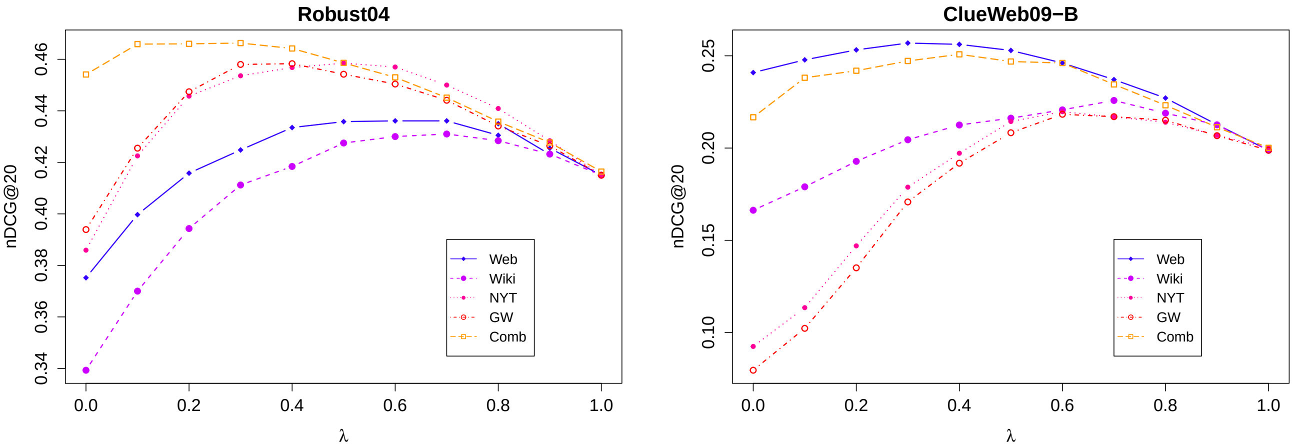

Figure 2: Retrieval performance (in nDCG@20) as a function of parameter λ.

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:

$$s_{comb}(Q,D) = \lambda \cdot P(Q|D) + (1-\lambda) \cdot \frac{1}{|\mathcal{S}|} \prod_{\sigma \in \mathcal{S}} \prod_{k \in \mathbb{T}^{\sigma}_{\hat{K},M}} \hat{\delta}_k \prod_{w \in \mathbb{W}_k} \hat{\phi}_{k,w} \cdot P(w|D) \qquad (7)$$

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 [14]. The results presented in Table 3 in the row **Comb** are not surprising and show support for the principles of *polyrepresentation* [20] and *intentional redundancy* [21] 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.

Finally, we explored 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 2, where high values of λ mean a high influence of the original query w.r.t 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.

4.5 Diversifying the result list

We just saw that different concepts learned from several sources of information can be complementary and very heterogeneous. We evaluate in this section at which point they can help retrieval diversity. Diversification is recent challenge in information retrieval which aims at providing the user with documents dealing with a broader range of subtopics when she issues an ambiguous query. To tackle this problem, the TREC Web track provides a major mean of evaluation through its diversity task [11]. By modeling concepts from various and broad sources of information, we also wanted to improve that document diversity. We evaluate in this section the performance of the four single resources as well as the full combination like in the previous section. Considering the Robust04 test collection do not have diversity-based relevance judgements we only perform evaluation with the ClueWeb09 and use the same 150 topics as before. We report results using the Web track’s official “intent-aware” metrics [1, 11] in Table 5.

	ERR-IA@20	α-nDCG@20
MRF	0.1717	0.2757
LCE	0.2046^{**}	0.2821
NYT	0.1699	0.2557
GW	0.1723	0.2668
Wiki	0.1749	0.2603
Web	$\mathbf{0.2172}^{***}$	0.3003^{*}
Comb	0.2081^{***}_{rr}	$\mathbf{0.3088}^{***}_{r}$

Table 5: Diversity evaluation of Latent Concept Modeling with respect to MRF and LCE (statistical tests are the same as in Table 3).

When using only one source of information for LCM, only the Web manages to significantly improve diversity over the baselines while the others achieves worst results than query likelihood. Despite results are not statistically significant for α-nDCG@20, the gain is of 8.9% over MRF and 6.4% over LCE. Again, linearly combining the concept models extracted for the four sources of information yields high results, close to the bests for ERR-IA@20 and statistically significant for α-nDCG@20. These results are consistent with other performances of our approach and confirm that it is robust and effective enough to outperform strong baselines in various retrieval tasks.

5. CONCLUSIONS

We presented in this paper an 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 performed consistently well over the two 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, which shows that our approach is robust enough to handle heterogeneous documents dealing with various topics to finally model concepts latent to the query.

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.

6. REFERENCES

[1] R. Agrawal, S. Gollapudi, A. Halverson, and S. Ieong. Diversifying search results. In *Proceedings of WSDM*, 2009.

[2] L. AlSumait, D. Barbará, and C. Domeniconi. On-line LDA: Adaptive Topic Models for Mining Text Streams with Applications to Topic Detection and Tracking. In *Proceedings of ICDM*, 2008.

[3] D. Andrzejewski and D. Buttler. Latent topic feedback for information retrieval. In *Proceedings of KDD*, 2011.

[4] R. Arun, V. Suresh, C. Veni Madhavan, and M. Narasimha Murthy. On Finding the Natural Number of Topics with Latent Dirichlet Allocation: Some Observations. In *Advances in Knowledge Discovery and Data Mining*, volume 6118 of *Lecture Notes in Computer Science*. 2010.

[5] J. Bai, J.-Y. Nie, G. Cao, and H. Bouchard. Using query contexts in information retrieval. In *Proceedings of SIGIR*, 2007.

[6] M. Bendersky, D. Metzler, and W. B. Croft. Parameterized concept weighting in verbose queries. In *Proceedings of SIGIR*, 2011.

[7] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3, 2003.

[8] J. Cao, T. Xia, J. Li, Y. Zhang, and S. Tang. A density-based method for adaptive LDA model selection. *Neurocomputing*, 72(7-9), 2009.

[9] B. Carterette and P. Chandar. Probabilistic models of ranking novel documents for faceted topic retrieval. In *Proceedings of CIKM*, 2009.

[10] Y. Chang, I. Ounis, and M. Kim. Query reformulation using automatically generated query concepts from a document space. *Information Processing & Management*, 42(2), 2006.

[11] C. L. A. Clarke, N. Craswell, I. Soboroff, and G. V. Cormack. Overview of the TREC 2010 Web Track. In *Proceedings of TREC*, 2010.

[12] S. Clinchant and E. Gaussier. Information-based models for ad hoc IR. In *Proceedings of SIGIR*, 2010.

[13] G. Cormack, M. Smucker, and C. Clarke. Efficient and effective spam filtering and re-ranking for large web datasets. *Information Retrieval*, 2011.

[14] F. Diaz and D. Metzler. Improving the estimation of relevance models using large external corpora. In *Proceedings of SIGIR*, 2006.

[15] O. Egozi, S. Markovitch, and E. Gabrilovich. Concept-Based Information Retrieval Using Explicit Semantic Analysis. *ACM Transactions on Information Systems*, 29(2), 2011.

[16] L. Finkelstein, E. Gabrilovich, Y. Matias, E. Rivlin, Z. Solan, G. Wolfman, and E. Ruppin. Placing search in context: the concept revisited. *ACM Transactions on Information Systems*, 20(1), 2002.

[17] D. Graff and C. Cieri. English Gigaword. *Philadelphia: Linguistic Data Consortium*, LDC2003T05, 2003.

[18] T. L. Griffiths and M. Steyvers. Finding scientific topics. *Proceedings of the National Academy of Sciences of the United States of America*, 101 Suppl, 2004.

[19] B. He and I. Ounis. Finding good feedback documents. In *Proceedings of CIKM*, 2009.

[20] P. Ingwersen. Polyrepresentation of information needs and semantic entities: elements of a cognitive theory for information retrieval interaction. In *Proceedings of SIGIR*, 1994.

[21] K. Jones. *Retrieving Information Or Answering Questions?* British Library annual research lecture. British Library Research and Development Department, 1990.

[22] M. Keikha, J. Seo, W. B. Croft, and F. Crestani. Predicting document effectiveness in pseudo relevance feedback. In *Proceedings of CIKM*, 2011.

[23] V. Lavrenko and W. B. Croft. Relevance based language models. In *Proceedings of SIGIR*, SIGIR '01, 2001.

[24] Y. Lu, Q. Mei, and C. Zhai. Investigating task performance of probabilistic topic models: an empirical study of PLSA and LDA. *Information Retrieval*, 2011.

[25] D. Metzler, Y. Bernstein, W. B. Croft, A. Moffat, and J. Zobel. Similarity measures for tracking information flow. In *Proceedings of CIKM*, 2005.

[26] D. Metzler and W. B. Croft. A Markov random field model for term dependencies. In *Proceedings of SIGIR*, 2005.

[27] D. Metzler and W. B. Croft. Latent concept expansion using markov random fields. In *Proceedings of SIGIR*, 2007.

[28] D. Newman, J. H. Lau, K. Grieser, and T. Baldwin. Automatic evaluation of topic coherence. In *Proceedings of HLT*, 2010.

[29] L. A. Park and K. Ramamohanarao. The Sensitivity of Latent Dirichlet Allocation for Information Retrieval. In *Proceedings of ECML PKDD*, 2009.

[30] S. E. Robertson and S. Walker. Some simple effective approximations to the 2-Poisson model for probabilistic weighted retrieval. In *Proceedings of SIGIR*, 1994.

[31] E. Sandhaus. The New York Times Annotated Corpus. *Philadelphia: Linguistic Data Consortium*, LDC2008T19, 2008.

[32] W. G. Stock. Concepts and semantic relations in information science. *Journal of the American Society for Information Science and Technology*, 61(10), 2010.

[33] T. Tao and C. Zhai. Regularized estimation of mixture models for robust pseudo-relevance feedback. In *Proceedings of SIGIR*, 2006.

[34] X. Wei and W. B. Croft. Lda-based document models for ad-hoc retrieval. In *Proceedings of SIGIR*, 2006.

[35] R. W. White, P. Bailey, and L. Chen. Predicting user interests from contextual information. In *Proceedings of SIGIR*, 2009.

[36] X. Yi and J. Allan. A Comparative Study of Utilizing Topic Models for Information Retrieval. In *Advances in Information Retrieval*, volume 5478 of *Lecture Notes in Computer Science*. 2009.

[37] C. Zhai and J. Lafferty. A study of smoothing methods for language models applied to information retrieval. *ACM Transactions on Information Systems*, 22(2), 2004.