Modeling topics in large text collections using search queries

Eduardo H. Ramírez and Ramon F. Brena

Tecnológico de Monterrey, México
{eduardo.ramirez,ramon.brena}@itesm.mx

Abstract. When analyzing collections using one of the well known probabilistic topic modeling methods, each resulting topic is represented as a probability distribution of terms. The assignment of topics to documents is also a probability value. This formalism has some limitations on computational complexity and in the human-understandability of results.
In this paper we propose an alternative unsupervised topic modeling approach, in which we model topics using automatically generated sets of keywords that are used as queries to an index of documents. By retrieving the documents relevant to those topical-queries we obtain overlapping clusters of semantically similar documents. Moreover, sets of keywords are useful as short human-readable descriptors of each topic. In order to find the topical-queries we present an approach consisting of generating candidate queries using signature-calculation heuristics and then evaluating candidates using an information-gain function defined as "semantic force". We present evidence to support the feasibility of this approach for semantically analyzing large collections.

1 Introduction

It is widely acknowledged that by using probabilistic methods to extract semantic information it is possible to improve access to information in collections in different application scenarios, such as retrieval [1] or collection browsing [2]. However, not every collection is subject to semantic analysis using existing methods because of their large scale, domain-specificity or human-effort-availability.

A central idea of state of the art approaches is the representation of topics as probability distributions of terms. This representation provides important theoretical advantages but also brings heavy computational limitations.

In this work, we focus on the problem of identifying topics in large text collections in a completely unsupervised way. In general, we aim to provide a comprehensive list of the topics that are discussed in the collection expressed in a human-readable format, the documents that belong to each topic, and the relative importance of the topics in the collection. Thus, in the aim of making topic mining feasible on large scale collections we propose an alternative method in which we represent topics as freely overlapping sets of documents, where each set can be retrieved using a boolean query, which we call the *"topical-query"*.

Instead of performing document-to-document comparisons, which may be very expensive in large collections, the proposed method focuses on generating candidates and searching for good topical-queries by their *"semantic force"*. The "semantic force" for a candidate query is computed by a real-valued function that given a query evaluates high when the retrieved documents are semantically similar. This approach enables us to take advantage of existing Web retrieval infrastructure, which is optimized for high-throughput.

We are specially interested in "cheap" semantic force functions that perform simple analysis on the retrieved results, such as counting them.

The main contribution of this paper is the formulation of a new topic identification problem that allows the abstract problem of topic modeling to be approached as a search over a candidate query space, evaluating each candidate query by its *"semantic force"*. We present also specific methods for leveraging those concepts into scalable algorithms. We show that the formulation makes the problem suitable to be implemented using parallel, massively scalable methods and provide an information-theoretical approximation to the *semantic force* function to show the feasibility of the approach.

The rest of the paper is organized as follows. In the next section we will discuss the abstract topic identification problem as it is approached today with an emphasis on its concrete instantiation as a Bayesian inference problem in Probabilistic topic modeling frameworks. In section 3 we will detail our core assumptions about the nature of topics and formulate a different version of the topic identification problem and then we will proceed to introduce our proposed method and its design concerns. In section 5 we will present experimental evidence to support the feasibility of the proposed approach and finally, in section 6 we will discuss the current status of the development, other related problems and the steps we plan to follow in order to properly situate this research into the current landscape of text-mining and its applications.

2 Related work

Broadly speaking, the abstract problem that concerns us is that of identifying the latent topic structure of a document collection in such a way that the former may be used to solve other computational problems.

The methods greatly vary in their computational strategy, which depends on the initial assumptions of what a topic is, how topics relate to documents and how topics relate to each other. However, all the referred methods, including ours, share the baseline assumption that corpus statistics contain enough information to produce useful results without requirements of expert knowledge. Consequently, solutions involving expert-crafted thesauri, such as Wordnet will not be discussed in this section.

Now, we will provide a quick summary of the major corpus-based unsupervised approaches to learning the semantic structure of collections. For comprehensive surveys on the field, please refer to [3].

2.1 Latent Semantic Indexing

Precursor works such as *Latent Semantic Indexing* [4] used a vectorial representation of words and documents, which could be used to arrange them in a space based on semantic similarity. An important motivation of this work was the retrieval of documents that did not contain the query terms, but were semantically similar to the ones that matched the user query. The number and structure of topics was only implicit in the proximity of terms in the concept space.

2.2 Probabilistic Topic Modeling

In the following years Hofmann [5] proposed a probabilistic version of LSI, namely *Probabilistic Latent Semantic Indexing* (PLSI). PLSI and all subsequent methods make the assumption that a document can be modeled as mixture of a number of hidden topics, and that those topics can be represented as probability distributions over words. Then, some sort of parameter estimation algorithm is applied to the observed data to estimate the parameters of the hidden topics. In the case of PLSI, the kind of estimation performed is a maximum likelihood estimation.

Later on, Blei et. al [6] proposed the *Latent Dirichlet Allocation* (LDA) as a generalization of PLSI. The key innovation in LDA was the introduction of fully generative semantics into the model formulation and thus allowing the problem to be treated by MCMC methods such as Gibbs sampling. In LDA each topic is represented as a multinomial distribution over words and each document is represented as a random mixture of topics, sampled from a Dirichlet distribution.

2.3 Discussion

Probabilistic topic models are a flexible and theoretically sound approach to learn topics in collections; however in relation to our area of concern which is (very) large scale collection analysis, the main limitation of the approach is the computational complexity which is heavily dependent on the number of topics in the model. Although there have been interesting proposals on how to improve the efficiency of the sampling [7] or performing distributed inference [8], achieving greater scalability for very large corpora seems to imply a trade-offs in the quality of the estimations by limiting the number of topics below the optimal values and reducing the sampling iterations before the convergence zone. In other cases, improving scalability may require the implementation of model-specific optimizations which result on a loss of generality of the models [9].

We conclude that there exists a need for an alternative approach to the abstract topic modeling problem, designed to work well in the scenarios where the following conditions are satisfied:

– It is considered good enough to know the top-k most probable words for a topic without a specific ordering or probability value estimation.

- It is enough to know which topics are discussed in the document, without requiring the assignment of specific probability values to each topic.
- The size of the collection is in the order of millions of documents of variable length, and the number of topics is unknown and presumably very large[1].
- The availability of state-of-the-art search engine technology, such as the ability to do parallel processing in map-reduce style and to serve thousands of queries per second using a distributed index.

3 Topic Identification problem

Based on the design concerns presented in section 2, we will introduce the *topic identification task* as an formulation of the abstract topic modeling problem. From the outset, independently from whatever topic representation could be used, a topic in a given corpus can be considered as a vector of quantities, one for each document, such that each of the quantities measures the relation of the given document to the topic. This notion of topic has the advantages of both being very general, but also being independent of any specific topic representation. Let us assume that those quantities are in the range 0 to 1. All particular topic representations, such as conditional probability vectors of words to appear in a document given it belongs to a topic [5], could be reduced to the above definition, once it is applied to a specific corpus.

Now, in our approach we are proposing to limit the quantities relating topics to documents to just 1 and 0, meaning respectively that the document belongs to the topic or not. Though this is a strong simplification, we stress that it is still useful, and that it could lead to much more efficient algorithms, as we show in the following.

Let $D = \{d_1, d_2...d_N\}$ be the corpus of documents of size N, $W = \{w_1, w_2...w_M\}$ the vocabulary of all M terms in the corpus and Q be the set of all boolean queries q_i that may be defined over the vocabulary W that will match at least one document from D.

A topic is $T_i \subset D$ is defined as a *set of documents* that may be retrieved using a boolean query $q_i \in Q$. Ideally, each T_i should contain documents that are semantically "similar" according to human criteria. In the "big picture", considering all possible topics in the complete corpus, the problem consists in finding an *optimal* set of *topical-queries* $Z^* = \{q_1, q_2...q_n\}$ and its associated document sets $T^* = \{T_1, T_2...T_n\}$, where each query q_i is a boolean expression over the terms in the vocabulary W and $T_i \neq \emptyset$.

Based on the definition of Q, a topical-query $q_i \in Q$ may be a conjunctive (AND) query: $q_i = \{w_{i,1} \wedge w_{i,2}... \wedge w_{i,k}\}$ or may be a disjunctive (OR) query composed of several conjunctive sub-queries, such that: $(q_i = \{q_1 \vee q_2\}) \Rightarrow (q_i = \{(w_{1,1} \wedge w_{1,2}... \wedge w_{1,n}) \vee (w_{2,1} \wedge w_{2,2}... \wedge w_{2,m})\})$.

A document is not constrained to belong to only one topic neither topics are required to be disjoints sets. Notice that we place no assumptions about

[1] Consider Wikipedia, if each article is considered a topic, the number of topics would be in the order of hundreds of thousands

the number of topics nor their structure, although the "topic-as-query" representation allows expressing hierarchical relations in a natural way, based on set containment. Besides being a computationally cheap way to represent topics, the chosen formalism allows to algorithmically perform manipulations such as merging, splitting and performing fast comparisons over topics using simple, well known metrics such as Jaccard's or Dice's similarity.

Let us call $F(q_i \in Q)$ the *"semantic force"* function we have mentioned, that evaluates the quality of a query and assigns a high value to the queries that retrieve semantically similar sets of documents.

Then, the set Z^* would correspond to the minimum set of topical-queries for which the value of the overall semantic force is maximized, where each $z_i \in Z$ represents a "coverage" of the corpus, that is, a set of queries $z_i = \{q_1, q_2 ... q_k\}$ that retrieve every element in D at least once.

$$Z^* = \arg\max_{z \in Z} \left\{ \frac{\sum_{q_i \in z} F(q_i)}{\sum n_d(q_i)} \right\} \tag{1}$$

The $n_d(q_i)$ function in equation 1 is a function that computes the number of documents retrieved by each topical-query, and it's required to compute an average of the total force by document, this is important considering that the same document may be retrieved several times and redundant coverages should be penalized. So, the global optimization criteria may be defined as *maximum semantic force per retrieved document*.

4 Proposed solution

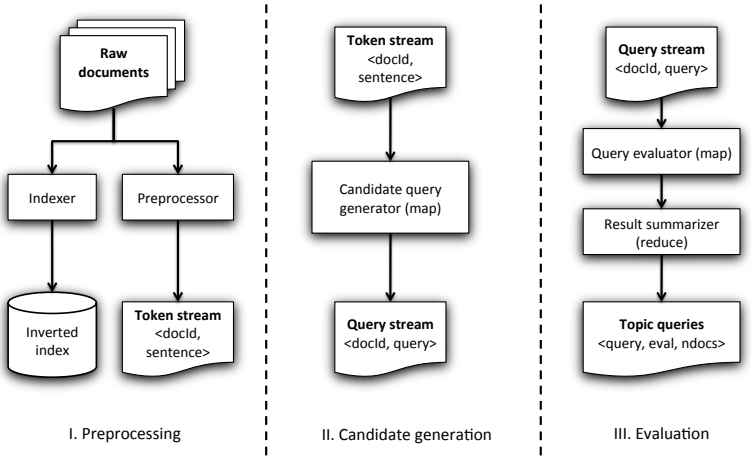


Fig. 1. Map/Reduce version of the topic identification algorithm

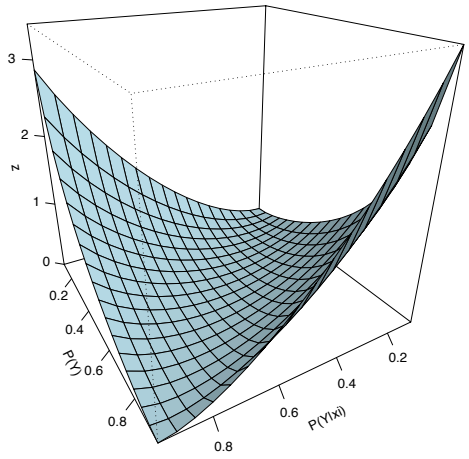


Fig. 2. KL-divergence as semantic force

We have implemented a fairly straightforward map/reduce algorithm that generates candidate queries using several heuristics and evaluated the queries using an information theoretical function that measures information gain. In this section we will provide the theoretical basis and technical decisions that support our approach.

4.1 Measuring the semantic force of queries

When approaching the topic identification task as a search problem over a query-space, the fundamental challenge that arises is that of producing a low-cost approximation to the ideal function $F(q_i \in Q)$. We propose a sound way to perform simple query alterations and measure the *amount of information* in result sets without performing extensive document-to-document comparisons.

Let q_i be a boolean query defined over a set of k words of the vocabulary and $W(q_i)$ the set of its terms, $W(q_i) = \{w_1, w_2..., w_k\}$ Now, we define two events for the experiment of selecting a random document of the corpus.

Let x be the event of retrieving a document with "any" of the terms of the query q_i. So, the probability of $P(x)$ is the probability of selecting a document retrieved by the query o_i, which is defined as a *disjunctive* (OR) query that matches any of the terms in $W(q_i)$, such that $o_i = \{w_1 \vee w_2... \vee w_k\}$.

Let y be the event of observing "all" the terms of the set $W(q_i)$. So, its probability $P(y)$ would be computed as the probability of selecting a document retrieved by a conjunctive query with all the terms of $W(q_i)$, like $a_i = \{w_1 \wedge w_2... \wedge w_k\}$.

Using these basic events, we may define a conditional event $y|x$ and $P(y|x)$ as the probability of observing *all* the query terms having observed *any* of them. So,

we propose to approximate the semantic force $F(q_i)$ computing the Kullback-Leibler divergence or *information gain* over this two events. KL-divergence is formally defined as:

$$K(P(Y|x)||P(Y)) = \sum_{y_j \in Y} P(y_j|x) log \frac{P(y_j|x)}{P(y_j)} \tag{2}$$

Where Y is the probability distribution defined over the events $\{y, \neg y\}$.

The KL-divergence can be directly interpreted as how much more certain we are about the fact that our randomly selected document will contain all the words of the query (y), given that it has any of them (x). The equation 2 measures the divergence about probability distributions, so we must take into consideration the complements of our events of interest $(\neg y)$, so by convenience we may express it as:

$$F(q_i) \approx P(y|x) log \frac{P(y|x)}{P(y)} + (1 - P(y|x)) log \frac{(1 - P(y|x))}{(1 - P(y))} \tag{3}$$

The proposed force function leverages the fact that semantically similar documents are more likely to have similar terms that those unrelated. So, if the query terms retrieve similar sets of documents whether executed as a conjunction or as a disjunction, we may infer that the query terms are semantically similar and thus, the retrieved documents also may be.

In figure 3 we present a plot of $K(P(Y|x)||P(Y))$ in the range $[0, 1]$. From the graph behavior we can observe that it will favor queries whose terms are unfrequent ($P(y) \approx 0$) and tend to co-occur with high probability ($P(Y|x) \approx 1$). In practice, this behavior require us to set a parameter $\alpha \in \mathbb{N}$ to assign a value of 0 to the topics below a minimum size of interest. If $\alpha = 1$, it will tend to favor document-specific topics.

4.2 Map/Reduce algorithm

The process overview is shown in figure 4 along with the key/value pairs that are computed in each phase of the process. Following the map/reduce style of design we rely as much as possible in stream-oriented operations, in this way, each step of the algorithm only needs to allocate enough memory to operate over a line of a stream at a time, that may be a sentence or a query.

Pre-processing The pre-processing phase takes the raw documents of the corpus as input and generates two outputs, that are required on the following phases. The first process that is performed is the document indexation into an inverted index. In order to evaluate the candidate queries, the index is frequently accessed, so, it is required to provide fast-index serving of boolean queries to the evaluation process. Stop-words were not removed during the indexing process, as they are important to the evaluation.

The second step in pre-processing involves the tokenization of the documents to create a stream of sentences. Each document is divided into sentences and a new line is produced for each sentence.

Candidate generation The candidate query generator takes as input the sentence stream and for every sentence it produces candidate queries. This implies the assumption that the existing topics in a document will affect the term co-occurrences probabilities at the sentence level. Also, it implies that the generated queries will capture some of the influence of proximity. Thus, we may say that the proposed approach does not assume a pure "bag-of-words" document model.

We have explored several methods to generate candidate queries, such as creating term combinations and computing the sentence n-grams. However, the method that so far performed better (in terms of number of candidate generated vs. evaluation) is based on a technique developed by Theobald et. al [10]. Interestingly, the technique was designed as a method to find duplicated documents by generating its semantic signatures or "spot-sigs". The method consists in splitting the sentences using a short list of stop-words as markers. In our implementation, we first split the list using the stop-words, and then generate additional 2-element combinations of the resulting units and keep all the queries that contain 2, 3 and 4 words. We probably generated an excessive number of candidates, however, at this point of the research we wanted to be sure that the subset of the query-space to analyze was large enough[2]. For this initial evaluation, all candidate queries are boolean conjunctive queries (AND) of the combined terms.

Candidate query evaluation and results summarization Candidate queries were evaluated using the semantic force function proposed in section 4.1. Each evaluation requires sending two queries to an index server. Depending on the size of the collection, the index may be on a single machine, distributed to the nodes in the map/reduce cluster or accessed remotely from an index serving cluster. Each deployment has different implications that need to be further researched.

The summarization process consists of determining the set of topical queries that would result in maximum semantic force per retrieved document. The first summarization step consists on selecting the maximum evaluated query for each document, in this way we can be sure that we have a complete coverage of the corpus. Then duplicates are removed, so we would have an initial topical-query set Z_0 of size bounded by $|Z_0| \leq |D|$, where each $q_i \in Z_0$ is a conjunctive query.

In the summarization phase, a set of specific conjunctive queries $q_1, q_2, ..q_k$ may be grouped into a more general disjunctive query $(q_1 \vee q_2 ... \vee q_k)$ in order to obtain the final topical query.

5 Experiments

We present the following experiments using the Reuters-21578 corpus to show the feasibility of the method. For our implementation the corpus was indexed using

[2] The Reuters-21578 corpus was splitted into 109,120 sentences from which 2,735,795 candidate queries were generated

Table 1. Performance for top-10 best identified topics

Topic	Relevant	Retrieved	Retrieved Relevant	Precision	Recall	Fscore
soybean	113	119	120	0.950	0.942	0.946
sorghum	24	24	35	1.000	0.686	0.814
wheat	287	401	307	0.716	0.935	0.811
sugar	119	128	184	0.930	0.647	0.763
silver	25	30	37	0.833	0.676	0.746
rapeseed	22	24	35	0.917	0.629	0.746
coffee	84	91	145	0.923	0.579	0.712
rubber	26	27	51	0.963	0.510	0.667
rye	1	1	2	1.000	0.500	0.667
grain	286	286	628	1.000	0.455	0.626

Apache Lucene with no stop-word filtering; in order to perform the sentence-splitting we are using the Lingua::EN::Sentence Perl module. We set $\alpha = 2$ to accept all potential discovered topics with more than two documents.

In order to show how the proposed function compares with "straw-man" approaches and establish a baseline performance, we compare the proposed KL-Divergence semantic-force with two simpler functions, defined as follows:

- Inverse Query Frequency (IQF). For a given query q_i, the inverse frequency of a query is defined as the logarithm of the inverse of the number of documents that matches q_i. $IQF(q_i) = log(N/n_d(q_i))$.
- TF-IQF. Analogous to tf-idf, this function is defined here as the number of sentences that generated q_i as a candidate within a document d_j, multiplied by the IQF as defined above. Formally, $TF\text{-}IQF(q_i, d_j) = n_s(q_i, d_j) * IQF(q_i)$.

The total execution time was 1 minute for indexing, 12 min for candidate generation and 30 for candidate evaluation, on an Intel 4x1Ghz, 4GB RAM machine running CentOs Linux. The unsummarized result set contained 9,027 distinct queries with 8,004 different terms.

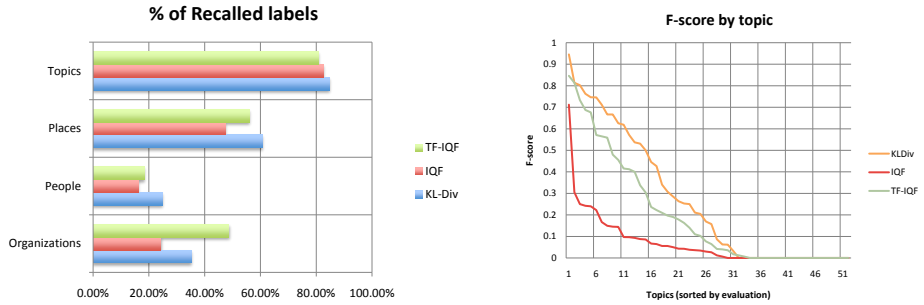
5.1 Topic label recall

In our first experiment we want to know to which extent the meaningful words in the corpus are captured in the topical queries result set, where each query is the maximum evaluated candidate for every document, without any further summarization.

The corpus contains 135 topic labels, from which we consider "recallable" the subset of 52 topics that exist in the vocabulary. In addition the corpus contains 175 place labels, 267 people names and 56 organizations, from which 37 exist in the vocabulary. Provisionally we consider these words as a human provided list of the most meaningful description of the corpus.

In figure 3(a) we present the initial performance result of the label recall task. Considering that the corpus was not prepared for the task, we consider the above 84% recall of topic labels an acceptable initial result. We hypothesize that

X



(a) Topic label recall task (b) Topic identification task

Fig. 3. Topic modeling performance

the recall performance decreases as a function of the complexity of terms (i.e. people and organization names are more complex than topics and places). In this benchmark the KL-Divergence function slightly outperforms the frequency based ones for all the concept classes except for organizations names.

5.2 Topic identification performance

Table 2. Average topic identification performance

	Precision	Recall	F-Score
KL-Div	0.767	0.375	0.441
IQF	0.870	0.075	0.124
TF-IQF	0.788	0.239	0.312

On the following experiments, we are concerned with measuring the similarity of the automatically discovered document sets with those produced by human labelers. For each topic label in the Reuters corpus we "executed" all the discovered topical queries that contained the label word and created a "retrieved" result set. A simple summarization was applied by chaining all the topical queries that contain each Reuters topic label as a disjunctive query.

The retrieved result is compared to the original labeled set and thus, recall, precision and f-measure may be computed in the usual way. The summarized results of the task are presented in figure 3(b) and table 2. For some topic labels no relevant documents were retrieved, so the results are splitted by total average and by the fraction of the topics for which at least 1 relevant document was retrieved, which are reported in the category "recalled".

From the "recalled" category we may observe that the KL-divergence semantic force function clearly outperforms the alternatives in terms of f-score being

the function that clearly provides best recall. This may be due to the fact that the KL-div function is the only one that considers the "disjunctive" or independent frequency of the query terms.

In absolute terms the results of the three functions were rather poor; however, keep in mind that we used the whole corpus to perform the task, while only approximately 50% of the documents are labeled with topics, so, the precision measurement may be "polluted" by unlabeled documents, that may or may not be relevant. Also, there exist cases in which the topic label does not exist in the document body, thus making the document not recallable at all.

6 Conclusion and future work

In this paper we approached the topic identification problem as the problem of finding sets of semantically similar documents by exploring the space of candidate queries that may be used to retrieve them.

We show that it is possible to find the topical-queries by performing simple query alterations and computing fast information-theoretic functions that only require to count the number of results to infer semantic similarity. We proposed a parallel framework that leverages Web retrieval infrastructure to perform such analysis and provided initial evidence about the feasibility of the approach.

Future work includes the design of a more complete benchmark for the topic identification task presented in this paper. We consider proposing an ad-hoc split of the Reuters corpus with clear recall/precision upper bounds. However, despite the corpus inherent difficulties, we consider that the presented experiments are valuable as they show that the technique has potential to match the human topic assignments.

Finally, we are working on improving the query summarization phase using a scalable unsupervised clustering algorithm. Following the information-theoretical approach, we believe that providing an accurate estimation of the number of topics based on objective criteria will be an important milestone for our work. In the mid-term our goal is to use the discovered topical queries to improve retrieval performance.

References

1. Li, R.M., Kaptein, R., Hiemstra, D., Kamps, J.: Exploring topic-based language models for effective web information retrieval. In Hoenkamp, E., de Cock, M., Hoste, V., eds.: Proceedings of the Dutch-Belgian Information Retrieval Workshop (DIR 2008), Maastricht, the Netherlands, Enschede, Neslia Paniculata (April 2008) 65–71
2. Blei, D., Lafferty, J.: A correlated topic model of science. **1**(1) (2007) 17–35
3. Griffiths, T.L., Steyvers, M., Tenenbaum, J.B.: Topics in semantic representation. Psychological Review **114** (2007) 211–244
4. Deerwester, S.C., Dumais, S.T., Landauer, T.K., Furnas, G.W., Harshman, R.A.: Indexing by latent semantic analysis. Journal of the American Society of Information Science **41**(6) (1990) 391–407

5. Hofmann, T.: Probabilistic latent semantic indexing. In: SIGIR '99: Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval, New York, NY, USA, ACM (1999) 50–57
6. Blei, D., Ng, A., Jordan, M.: Latent dirichlet allocation. Journal of Machine Learning Research **3** (2003) 993–1022
7. Porteous, I., Newman, D., Ihler, A., Asuncion, A., Smyth, P., Welling, M.: Fast collapsed gibbs sampling for latent dirichlet allocation. In: KDD '08: Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining, New York, NY, USA, ACM (2008) 569–577
8. Newman, D., Asuncion, A., Smyth, P., Welling, M.: Distributed inference for latent dirichlet allocation. In: Advances in Neural Information Processing Systems. Volume 20. (2007)
9. Wei, X., Croft, W.B.: Lda-based document models for ad-hoc retrieval. In: SIGIR '06: Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval, New York, NY, USA, ACM (2006) 178–185
10. Theobald, M., Siddharth, J., Paepcke, A.: Spotsigs: robust and efficient near duplicate detection in large web collections. In: SIGIR '08: Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval, New York, NY, USA, ACM (2008) 563–570