Multi-objective topic modeling for exploratory search in tech news

Anastasia Ianina[1], Lev Golitsyn[2], and Konstantin Vorontsov[3]

[1] Moscow Institute of Physics and Technology
yanina@phystech.edu
[2] Integrated Systems
lvgolitsyn@gmail.ru
[3] Moscow Institute of Physics and Technology
vokov@forecsys.ru

Abstract. Exploratory search is a paradigm of information retrieval, in which the user's intention is to learn the subject domain better. To do this the user repeats "query–browse–refine" interactions with the search engine many times. We consider typical exploratory search tasks formulated by long text queries. People usually solve such a task in about half an hour and find dozens of documents using conventional search facilities iteratively. The goal of this paper is to reduce the time-consuming multi-step process to one step without impairing the quality of the search. Probabilistic topic modeling is a suitable text mining technique to retrieve documents, which are semantically relevant to a long text query. We use the additive regularization of topic models (ARTM) to build a model that meets multiple objectives. The model should have sparse, diverse and interpretable topics. Also, it should incorporate metadata and multimodal data such as n-grams, authors, tags and categories. Balancing the regularization criteria is an important issue for ARTM. We tackle this problem with coordinate-wise optimization technique, which chooses the regularization trajectory automatically. We use the parallel online implementation of ARTM from the open source library BigARTM. Our evaluation technique is based on crowdsourcing and includes two tasks for assessors: the manual exploratory search and the explicit relevance feedback. Experiments on two popular tech news media show that our topic-based exploratory search outperforms assessors as well as simple baselines, achieving precision and recall of about 85–92%.

Keywords: information retrieval, exploratory search, relevance feedback, topic modeling, additive regularization for topic modeling, ARTM, BigARTM

1 Introduction

Exploratory search is a relatively new paradigm in information retrieval. It aims to satisfy advanced information needs of people for education, self-education, knowledge acquisition and discovery [11,21]. Potential users of exploratory search are students, teachers, researchers and professionals. In knowledge society, the information needs of users increase constantly and become more and more complicated. This leads to the emergence of new search paradigms and tools.

In exploratory search, the user may not be familiar with the terminology and may assume that there are many correct answers. The user's search intent may be just learning the basics of the subject domain and defining the most important topics within it. In such cases it is difficult or even impossible to formulate an exact short query. The user of a conventional search system has to enter many queries iteratively, gradually learning the terminology and refining his or her knowledge and intentions. The iterative "query–browse–refine" process [21] may require a lot of time and experience. The alternative way is to indicate a broad search direction by a long text query, such as a whole document, a set of copy-pasted text fragments, or a document folder, and give the user a set of semantically similar documents. There are two obstacles along this way. The first one is in elaborating a semantic similarity measure appropriate for the purposes of exploratory search. The second one is in evaluating both precision and recall, which is a difficult task for human assessors. In order to address these challenges, we propose a topic-based approach to exploratory search and a three-stage model evaluation and selection technique based on crowdsourcing.

Topic modeling is often used for searching semantically similar documents [1, 20, 22] and has become more popular in exploratory search community in recent years [8, 12, 13, 15]. The *probabilistic topic model* reveals the latent thematic structure of a text collection. It determines each topic as a discrete probability distribution over words and then represents each document by a discrete probability distribution over topics [5,6,9]. The conventional full text search is usually based on the inverted index and looks for documents, which contain all the words from the query [10]. So, if the query is long, it's most likely that nothing will be found. Topic-based search overcomes this problem by using compact topic vector representations for the query and documents instead of their bag-of-words representations. This way, one can use the same mechanisms of indexing and ranking for searching topically similar documents, it's just that now topics take the place of words.

To be used in the exploratory search system, the topic model has to meet multiple requirements. Topics should be significantly different and well interpretable to capture semantics appropriately. Vector representations of documents should be highly sparse to make the inverted index as compressed as possible. The model should take into account the modalities of authors, time stamps, categories, tags, named entities etc. to get the most out of the available meta-information. We use a multi-objective approach called *additive regularization of topic models* (ARTM) [17] to satisfy all these requirements. ARTM learns models with desired properties by maximizing a weighted sum of the log-likelihood and additional regularization criteria. We use an effective parallel implementation of the expectation-maximization (EM) algorithm from open source project BigARTM.org [7]. Our experiments show that the combination of the above requirements in a form of regularization criteria significantly improves not only the model itself, but also precision and recall of the exploratory search.

Two popular tech news media are used for the evaluation: techcrunch.com in English and habrahabr.ru in Russian. Our evaluation technique consists of three stages. At the first stage we ask assessors to find the documents relevant to the long-text queries using any search utilities of their choice. At the second stage we ask assessors to give *explicit relevance feedback* [4] for the topic-based search results on the same queries. At the third stage we join for each query all sets of relevant documents found at the previous stages. These enriched assessor data enables us to estimate precision and recall for new models. In addition, we get the opportunity to compare and select models without asking assessors.

Assessors spend about 30 minutes on average per a query. For this reason we afford to collect a limited amount of assessor data sufficient for model validation and selection. Learning the supervised topic model would require much more assessor data. However, this is not necessary, since the multi-objective unsupervised topic model already provides a high quality exploratory search.

The paper is organized as follows. In section 2 we introduce the ARTM framework and describe the strategy of choosing regularization coefficients. In section 3 we describe the evaluation technique for the topic-based exploratory search. In section 4 we reports the experimental results of comparing topic-based search with baselines. In section 5 we use assessor data for model selection. In section 6 we conclude that topic-based exploratory search is much faster than assessors' iterative search, having better recall and comparable precision.

2 Probabilistic topic modeling and additive regularization

Let us denote a finite set (collection) of texts by D, a finite set of topics by T, and a finite set of modalities by M. Here are some examples of modalities: words, bigrams, tags, categories, authors, etc. Each modality $m \in M$ has a finite set (dictionary) of tokens W_m. Each document $d \in D$ is a sequence of n_d tokens from $W = \bigcup_w W_m$. We accept the bag-of-words hypothesis and take into account how many times n_{dw} the token w appears in the document d.

Given the $(n_{dw})_{D\times W_m}$ matrix, a probabilistic topic model finds its approximate matrix factorization by $\varPhi_m = (\phi^m_{wt})_{W_m\times T}$ matrix of token probabilities for the topics and $\varTheta = (\theta_{td})_{T\times D}$ matrix of topic probabilities for the documents:

$$\frac{n_{dw}}{n_d} \approx p(w\,|\,d) = \sum_{t\in T} p(w\,|\,t)\,p(t\,|\,d) = \sum_{t\in T} \phi_{wt}\theta_{td},$$

where $|T|$ is a user-defined number of topics in the model.

Usually, the problem of matrix factorization has infinitely many solutions. Additive regularization [17, 19] narrows the set of solutions by maximizing the weighted sum of modality log-likelihoods and regularizers $R_i(\varPhi, \varTheta)$:

$$\sum_m \tau_m \sum_{d\in D} \sum_{w\in W_m} n_{dw} \ln \sum_{t\in T} \phi_{wt}\theta_{td} + \sum_{i=1}^{r} \tau_i R_i(\varPhi, \varTheta) \to \max_{\varPhi,\varTheta}$$

under non-negativity and normalization constraints for all columns of $\varPhi_m$ and $\varTheta$ matrixes. This optimization problem can be solved using the EM-algorithm [17]. Many topic models can be considered as special cases of additive regularization (ARTM) with appropriate choice of regularizers [16, 17]. Regularization coefficients τ_m and τ_i are usually chosen empirically.

Probabilistic Latent Semantic Analysis (PLSA) [9] corresponds to the absence of regularization, $R(\varPhi, \varTheta) = 0$.

Latent Dirichlet Allocation (LDA) [6] corresponds to the *smoothing* regularizer, which minimizes the cross-entropy between columns $\boldsymbol{\phi}_t$ and a fixed distribution $\boldsymbol{\beta} = (\beta_w \colon w \in W)$ as well as the cross-entropy between columns $\boldsymbol{\theta}_d$ and a fixed distribution $\boldsymbol{\alpha} = (\alpha_t \colon t \in T)$:

$$R(\varPhi, \varTheta) = \beta_0 \sum_{t \in T} \sum_{w \in W} \beta_w \ln \phi_{wt} + \alpha_0 \sum_{d \in D} \sum_{t \in T} \alpha_t \ln \theta_{td}, \tag{1}$$

where positive vectors $\beta_0 \boldsymbol{\beta}$ and $\alpha_0 \boldsymbol{\alpha}$ are interpreted as hyperparameters of Dirichlet prior distributions in the Bayesian topic modeling framework. Scalars β_0 and α_0 are interpreted as regularization coefficients in the ARTM framework. Choosing uniform distributions for $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ corresponds to symmetric Dirichlet priors, which are often used in experiments with the LDA model.

The *sparsing regularizer* has the same form as in (1), but differs in that the coefficients β_0 and α_0 are negative [17]. Sparsing maximizes the cross-entropy enforcing columns $\boldsymbol{\phi}_t$ and $\boldsymbol{\theta}_d$ to be as far as possible from distributions $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ respectively. This regularizer can not be interpreted in terms of Dirichlet priors.

The *decorrelation regularizer* makes topics as different as possible by minimizing the sum of covariances between topic vectors $\boldsymbol{\phi}_t$:

$$R(\varPhi) = -\sum_{t,s \in T} \sum_{w \in W} \phi_{wt} \phi_{ws}.$$

Diversifying the term distributions of topics is known to make the resulting topics more interpretable [14]. Also, this regularizer stimulates sparsity and tends to group stop-words and common words into separate topics.

The combination of three regularizers above improves the interpretability of topics [2, 3, 17, 18]. In our experiments we also use the combination of three regularizers: decorrelation of term distributions in topics with the coefficient τ, sparsing topic distributions in documents with the coefficient α, smoothing term distributions in topics with the coefficient β.

We subsequently add regularizers to the model following empirical recommendations from [17]: decorrelation goes first, then smoothing and sparsing. Generally, the sequential strategy enables a regularizer to prepare data for the following ones or to compensate side-effects of the previous ones. In our case, decorrelation rotates topic vectors $\boldsymbol{\phi}_t$ to make them more distinct, $\varPhi$-smoothing compensates for the excessive sparsing after decorrelation, and $\varTheta$-sparsing nullifies insignificant probabilities when the process is close to convergence.

For each regularizer we choose its regularization coefficient from a grid of values using multiple criteria. In our experiments we use the following criteria: perplexity, $\varPhi$-sparsity, and $\varTheta$-sparsity. We perform 8 iterations of the EM-algorithm

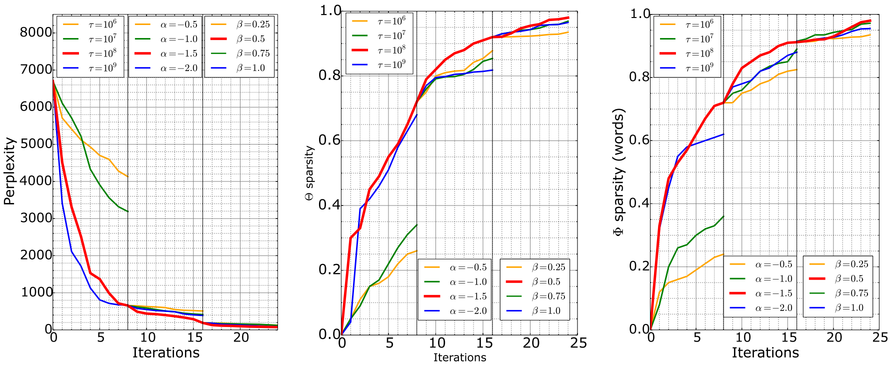

Fig. 1: Choosing regularization coefficients on Habrahabr collection. Perplexity, Θ and Φ sparsity depending on iteration count.

for each value of each coefficient. Thus, every model is trained along its regularization trajectory, which consists of $3 \cdot 8 = 24$ iterations. From all regularization trajectories we choose the one that yields an improvement in at least one of the criteria without a significant impairment in the others. So, our technique for tuning the regularization coefficients is a particular case of coordinate-wise optimization with grid search along each coordinate. An example of a regularization trajectory is shown in Fig. 1 for the Habrahabr collection.

The optimization of the regularization trajectory is fully automated for further model selection. In section 5 it will be used for the selection of the number of topics, the set of modalities, and the semantic similarity measure.

3 Topic-based exploratory search

An exploratory search query q is a long text, so we learn its topic vector θ_q in the same way as it was done for the documents in the collection. Next, among topic vectors of documents θ_d, we find k documents closest to the query and return them as a search result.

Similarity between queries and documents can be measured using cosine similarity, Euclidean distance, Manhattan distance, Hellinger distance, Kullback–Leibler divergence, or others. In section 5, we will empirically compare the search quality they yield.

For evaluating the results of topic-based exploratory search we simulate situations that analysts might encounter in practice when preparing reviews or digests of technical news. We form a set of long thematically focused text queries relevant to the collection (Fig. 2). On average, a query consists of roughly a single A4 page of text (Fig. 3). Each query is composed of fragments copy-pasted from texts both inside and outside the collection. The query should be sufficiently complete, so as to minimize discrepancies in its interpretation by different as-

3D-printers
AB-testing in huge IT corporations
Algorithms for searching a minimal spanning tree
New Amazon Kindle products
Apple product presentations
Best-known Y Combinator projects
CERN-cluster
Communication within employees in large companies
Cryptosystems with public keys
Daily planners (mobile applications)
Internet of things
Hadoop MapReduce
Healthcare devices
How to write a good CV
LogService (Facebook system for storing logs)
Main educational sources for data scientists
MIT MediaLab research
Online education
Self-driving cars
Seq2seq neural networks

Fig. 2: Examples of titles for 20 exploratory search queries

Title: SpaceX Falcon Launch

SpaceX has successfully launched a Falcon 9 to orbit during its BulgariaSat-1 mission Friday. The launch reused a first stage booster first employed during an Iridium Communications mission in January of this year, after that Falcon 9 first stage was recovered and refurbished.
Elon Musk has shared a new animation created by SpaceX to demonstrate the planned launch process for its Falcon Heavy rocket, which it hopes to test fly for the first time this coming November. The animation depicts launch of the three-booster heavy rocket, separation of the first and second stages, and the return flight and landing of the three booster cores used to get the rocket to space.
SpaceX has completed the other key ingredient of its historic flight, recovering its Falcon 9 rocket via its floating drone barge. This is a huge accomplishment because it already did this once before – with the same rocket, on the same barge, when it landed last year following a successful launch during a resupply mission to the International Space Station.
The recovery of the Falcon 9 means that not only did SpaceX reuse its rocket with this launch – it can also potentially use it again, after more stress testing and evaluation.
Its hard to underscore the significance of this milestone, but theres still ample work to do: SpaceXs goal is to eventually be able to relaunch rockets within the same day, which is obviously a feat on a different scale.

Fig. 3: An example of an exploratory search query

sessors. On the other hand, the query should be short enough for an assessor to understand its essence quickly.

For each query we ask an assessor to perform two sequential tasks.

In the first task, an assessor is asked to find within the collection as many documents relevant to the query as possible. The assessor may use any search tools available: a built-in search line, hyperlinks, tags or categories, a conventional search system such as Google, Bing, Yandex etc. This task is rather creative, usually taking a person about half an hour to complete. The time taken to process a query is recorded.

In the second task, the assessor is asked to look through the list of documents retrieved by the topic-based search for the same query and mark each document as relevant or irrelevant. Thus, we get the explicit relevance feedback for the topic-based search.

Each query is processed by 3 assessors to reduce the variance of the result and to find more relevant documents.

For each query we measure the quality by two metrics: Precision@k and Recall@k. Precision@k is the fraction of relevant documents among the first k documents found. Recall@k is the fraction of found relevant documents among all the relevant documents. We take the average Precision@k and Recall@k over all queries and over all assessors to evaluate the topic search quality.

The calculation of Recall requires knowing the set of all relevant documents for each query. We approximate this set by joining all the documents that were found by all assessors during the first task and all the documents that were found by topic-based search and confirmed by the majority of assessors as relevant during the second task. We also expanded the sets of relevant documents with the search results returned by baseline algorithms. However, this expansion has given very few relevant documents. From here we conclude that the obtained sets of relevant documents are close to being complete, and that they are suitable for comparing the search algorithms.

4 Experiments with topic-based search

Datasets. The experiments were conducted on two tech news collections — TechCrunch.com in English and Habrahabr.ru in Russian. Text pre-processing included deleting punctuation, bringing the upper case letters down to the lower case and lemmatizing using the morphological analyzer pymorphy2.

The TechCrunch collection consists of 759324 articles. Articles contain tokens of four modalities: 11523 word unigrams, 1.2 mln. bigrams (the tail of rare bigrams was deleted), 605 authors and 184 categories.

The Habrahabr collection consists of 175143 articles. Articles contain tokens of six modalities: 10552 word unigrams, 742000 word bigrams, 524 authors, 10000 commentators (authors of comments to the articles), 2546 tags, 123 hubs (categories). We exclude 5 percent of the most frequent words in the collection.

Topic-based search vs. assessors. We applied the evaluation method described above to the Habrahabr and Techcrunch collections. For Habrahabr we constructed 100 queries by copying and merging fragments of text taken from sources outside Habrahabr such as other IT-oriented blogs, posts from stackoverflow.com, articles from ixbt.com, etc. The length of a query ranges from 93 to 455 words with the average of 262 words.

The experiment results for the Habrahabr collection are presented in Fig. 4. The points on the plot correspond to queries. We compare precision and recall of the search performed by the assessors with the topic-based search for the best of our models. On average, precision is a bit higher for assessors' search, whilst recall is higher for the topic-based search. The highest recall we got for the topic-based search is 1.0 for 26 queries out of 100. From the right chart in Fig. 4 it can be seen that there is no obvious dependence between the time spent by an assessor and the quality of the search. On average, it took assessors about 30 minutes to process a single query. The number of relevant articles ranges from 5 to 55, the average being 25.

The experiment for the TechCrunch collection is presented in Fig. 5. There were 100 queries and each of them was processed by 3 assessors. The length of the query ranges from 75 to 392 words, the average being 195 words. The average number of articles found by assessors per query is 32.

Thus, topic-based exploratory search obtains higher recall and produces the results significantly faster than human assessors. In some cases, topic-based

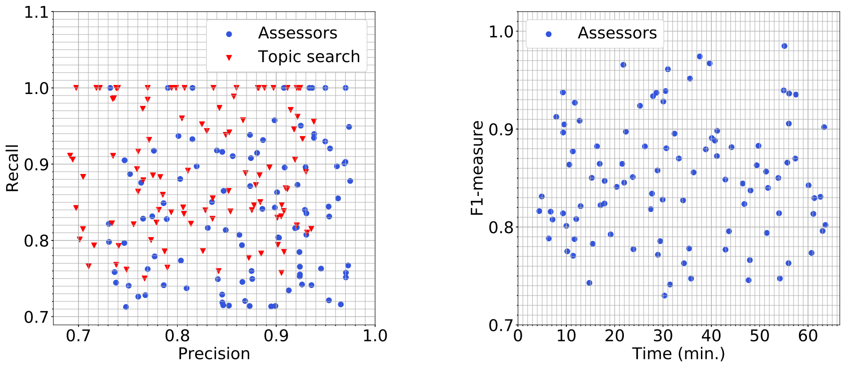

Fig. 4: The quality of assessors' and topic-based exploratory search (Habrahabr)

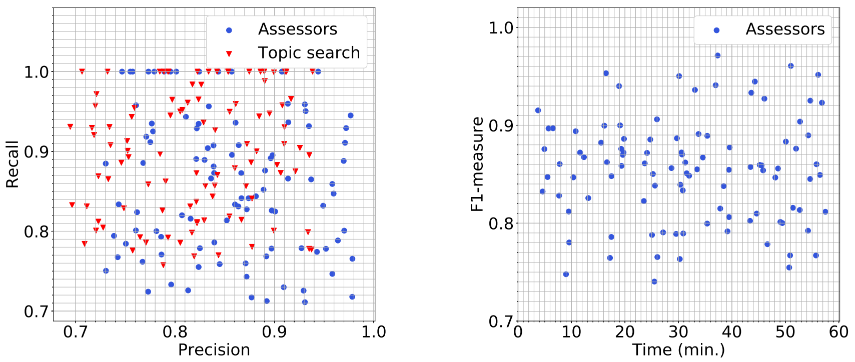

Fig. 5: The quality of assessors' and topic-based exploratory search (TechCrunch)

search finds relevant documents that all three assessors have missed during the first task.

The significance of the difference in precision/recall between assessors' search and topic-based search was tested using the Wilcoxon signed-rank test. For all tests the p-value was less than 0.01. From here we conclude that the dataset of 100 queries is sufficient to compare the search quality.

Topic-based search vs. baselines. We use a simple but strong full-text TF-IDF search as a first baseline. We apply lemmatization to Russian texts and stemming to English texts. Then we get TF-IDF vectors from documents and queries using a simple vectorizer from the sklearn library. As a search result, we return those k documents that have TF-IDF vectors closest to the query. The TF-IDF search is a strong competitor for the topic-based search because it uses full information from word-document frequency matrix, whilst the topic-based search uses the

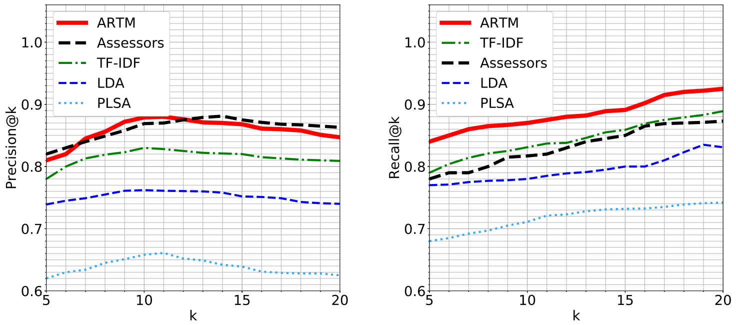

Fig. 6: Comparison of assessors' and topic-based search with regularization (ARTM) and baselines (TF-IDF, PLSA, LDA) for Habrahabr

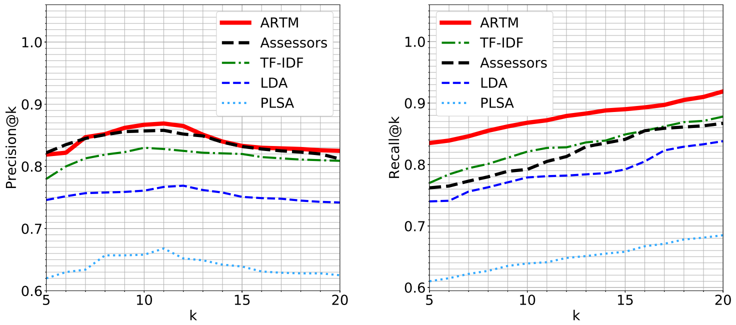

Fig. 7: Comparison of assessors' and topic-based search with regularization (ARTM) and baselines (TF-IDF, PLSA, LDA) for TechCrunch

low-rank approximation of this matrix. To make the baseline even stronger we take into consideration not only words, but also tags and categories. According to Fig. 6 and Fig. 7, topic-based search gives better results in terms of precision and recall than the TF-IDF search. This fact confirms that the topic model gives a rich semantic representation of documents and queries.

Another advantage of the topic-based search in comparison to TF-IDF search is that the low-dimensional sparse topical representation of documents can be converted into a highly compressed inverted index. Hence, an effective topic-based search engine can be implemented at low cost.

Also we introduce two additional baselines based on PLSA and LDA topic models respectively. Experiments show that they both perform worse than the ARTM-based search, see Fig. 6 and Fig. 7.

The Wilcoxon signed-rank test test has confirmed that the differences between our search and the baselines are significant: p-values were less than 0.0004 in 48 tests for Precision@k, Recall@k, $k \in \{5, 10, 15, 20\}$, all three baselines, and both collections.

Table 1: Topic-based search with different sets of regularizers: Decorrelation, Θ-sparsing, Φ-smoothing

	Habrahabr				**TechCrunch**			
	no reg	D	DΘ	D$\Theta\Phi$	no reg	D	DΘ	D$\Theta\Phi$
Pr@5	0.628	0.748	0.771	**0.810**	0.652	0.775	0.779	**0.819**
Pr@10	0.653	0.776	0.812	**0.879**	0.679	0.787	0.819	**0.867**
Pr@15	0.642	0.765	0.792	**0.868**	0.669	0.773	0.798	**0.833**
Pr@20	0.643	0.759	0.783	**0.847**	0.673	0.777	0.792	**0.825**
R@5	0.692	0.784	0.805	**0.840**	0.673	0.812	0.812	**0.835**
R@10	0.714	0.814	0.834	**0.870**	0.685	0.821	0.845	**0.868**
R@15	0.725	0.835	0.867	**0.891**	0.712	0.859	0.869	**0.890**
R@20	0.735	0.862	0.891	**0.925**	0.723	0.882	0.895	**0.919**

Table 2: Topic-based search with different similarity measures: Euclidean, Cosine, Manhattan, Hellinger, Kullback-Leibler

	Habrahabr					**TechCrunch**				
	Eu	**cos**	Ma	He	KL	Eu	**cos**	Ma	He	KL
Pr@5	0.612	**0.810**	0.682	0.709	0.721	0.635	**0.819**	0.673	0.732	0.715
Pr@10	0.657	**0.879**	0.697	0.735	0.749	0.665	**0.867**	0.683	0.752	0.732
Pr@15	0.627	**0.868**	0.635	0.727	0.711	0.643	**0.833**	0.642	0.742	0.724
Pr@20	0.619	**0.847**	0.627	0.728	0.707	0.638	**0.825**	0.638	0.729	0.708
R@5	0.672	**0.840**	0.692	0.721	0.803	0.658	**0.835**	0.669	0.733	0.775
R@10	0.682	**0.870**	0.707	0.775	0.856	0.671	**0.868**	0.682	0.753	0.787
R@15	0.705	**0.891**	0.725	0.791	0.878	0.715	**0.890**	0.708	0.785	0.809
R@20	0.703	**0.925**	0.732	0.812	0.888	0.712	**0.919**	0.715	0.808	0.812

The importance of regularizers. To show that each regularizer is important and significantly improves the search quality we carry out one more experiment. Table 1 shows that the decorrelation regularizer contributes the most to the search quality, but the other regularizers are also necessary. The model with no regularization gives the worst result.

5 Model parameters optimization

Sets of relevant documents found by assessors for every query allow us to evaluate new topic models or new search algorithms without any additional assessment. Below we describe three experiments in which three hyperparameters were selected alternately (the similarity measure, the set of modalities, and the number of topics), while the other two were fixed to be optimal.

Table 2 shows that cosine similarity is the best similarity measure between query and document topic vectors. The topic model used in this experiment has the optimal number of topics and the full set of modalities.

Table 3 shows that the use of all modalities together improves both recall and precision of the search. Terms and tags contribute the most. Models with

Table 3: Topic-based search using different modalities
Habrahabr: Assessors, Words, Bigrams, Comments, Tags, Hubs, Authors
TechCrunch: Assessors, Words, Bigrams, Authors, Categories

	Habrahabr						**TechCrunch**					
	As	W	C	WB	WBTH	**All**	As	W	C	WB	WBC	**All**
Pr@5	0.821	0.612	0.549	0.654	0.737	**0.810**	0.822	0.711	0.557	0.767	0.808	**0.819**
Pr@10	0.869	0.635	0.568	0.701	0.752	**0.879**	0.851	0.721	0.581	0.783	0.818	**0.867**
Pr@15	0.875	0.625	0.532	0.685	0.682	**0.868**	0.835	0.733	0.594	0.793	0.833	**0.833**
Pr@20	0.863	0.616	0.533	0.682	0.687	**0.847**	0.813	0.727	0.566	0.772	0.822	**0.825**
R@5	0.780	0.722	0.636	0.797	0.827	**0.840**	0.762	0.752	0.657	0.775	0.825	**0.835**
R@10	0.817	0.744	0.648	0.812	0.875	**0.870**	0.792	0.776	0.669	0.808	0.855	**0.868**
R@15	0.850	0.778	0.677	0.842	0.893	**0.891**	0.835	0.782	0.684	0.825	0.877	**0.890**
R@20	0.873	0.803	0.685	0.852	0.898	**0.925**	0.867	0.825	0.702	0.837	0.901	**0.919**

Table 4: Topic-based search using a different number of topics

	Habrahabr						**TechCrunch**					
	As	100	150	**200**	250	400	As	350	400	450	**475**	500
Pr@5	0.821	0.662	0.721	**0.810**	0.761	0.693	0.822	0.653	0.725	0.752	**0.819**	0.777
Pr@10	0.869	0.761	0.812	**0.879**	0.825	0.673	0.851	0.663	0.732	0.762	**0.867**	0.811
Pr@15	0.875	0.733	0.795	**0.868**	0.791	0.651	0.835	0.682	0.743	0.787	**0.833**	0.793
Pr@20	0.863	0.724	0.795	**0.847**	0.792	0.642	0.813	0.650	0.743	0.773	**0.825**	0.793
R@5	0.780	0.732	0.807	**0.840**	0.821	0.721	0.762	0.731	0.762	0.793	**0.835**	0.817
R@10	0.817	0.771	0.843	**0.870**	0.851	0.751	0.792	0.763	0.793	0.812	**0.868**	0.855
R@15	0.850	0.824	0.895	**0.891**	0.871	0.773	0.835	0.782	0.807	0.855	**0.890**	0.882
R@20	0.873	0.857	0.905	**0.925**	0.892	0.771	0.867	0.792	0.823	0.862	**0.919**	0.903

only one modality show the worst results. All the models used in this experiment have the optimal number of topics.

Table 4 shows that an optimal number of topics $|T|$ for the model having the full set of modalities equals 200 for Habrahabr, 475 for TechCrunch.

The whole set of experiments shows that the optimal number of topics stays the same for all similarity measures, and the optimal set of modalities stays the same for all similarity measures and all values of $|T|$.

6 Conclusions

In this paper, we propose an additively regularized topic model for exploratory search of relevant documents by long text queries. We show that the combination of decorrelation, sparsing and smoothing regularizers originally designed to improve the model interpretability also improves the search quality. We also confirm that the model should incorporate all available meta-data and modalities, such as bigrams, authors, tags, and categories.

For evaluating both precision and recall of the search we use an empirical technique based on human assessments. We achieve high quality results on realistic tasks of exploratory search in tech news. It seems that this level of quality would be enough for applications, such as automation of writing reviews and information consolidation. The topic-based search instantly performs the work that people typically complete in about 30 minutes. Another advantage of topic-based search over conventional full-text search is in reduction of the size of the inverted index, which enables an effective and low-cost implementation.

Acknowledgements. The work was supported by the Government of the Russian Federation (agreement 05.Y09.21.0018), the Ministry of Education and Science of the Russian Federation (project RFMEFI57915X0117), and the Russian Foundation for Basic Research (grants 17-07-01536, 16-37-00498). We are thankful to Maria Veretennikova for her help.

References

[1] Andrzejewski, D., Buttler, D.: Latent topic feedback for information retrieval. In: Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 600–608. KDD '11 (2011)

[2] Apishev, M., Koltcov, S., Koltsova, O., Nikolenko, S., Vorontsov, K.: Additive regularization for topic modeling in sociological studies of user-generated text content. In: MICAI 2016, 15th Mexican International Conference on Artificial Intelligence. vol. 10061, pp. 166–181. Springer, Lecture Notes in Artificial Intelligence (2016)

[3] Apishev, M., Koltcov, S., Koltsova, O., Nikolenko, S., Vorontsov, K.: Mining ethnic content online with additively regularized topic models. Computacion y Sistemas 20(3), 387–403 (2016)

[4] Baeza-Yates, R., Ribeiro-Neto, B.: Modern Information Retrieval: The Concepts and Technology behind Search (2nd Edition) (ACM Press Books). Addison-Wesley Professional (2011)

[5] Blei, D.M.: Probabilistic topic models. Communications of the ACM 55(4), 77–84 (2012)

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

[7] Frei, O., Apishev, M.: Parallel non-blocking deterministic algorithm for online topic modeling. In: AIST'2016, Analysis of Images, Social networks and Texts. vol. 661, pp. 132–144. Springer International Publishing Switzerland, Communications in Computer and Information Science (CCIS) (2016)

[8] Grant, C.E., George, C.P., Kanjilal, V., Nirkhiwale, S., Wilson, J.N., Wang, D.Z.: A topic-based search, visualization, and exploration system. In: FLAIRS Conference. pp. 43–48. AAAI Press (2015)

[9] Hofmann, T.: Probabilistic latent semantic indexing. In: Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval. pp. 50–57. ACM, New York, NY, USA (1999)

[10] Manning, C.D., Raghavan, P., Schütze, H.: Introduction to Information Retrieval. Cambridge University Press, New York, NY, USA (2008)

[11] Marchionini, G.: Exploratory search: From finding to understanding. Commun. ACM 49(4), 41–46 (2006)

[12] Rönnqvist, S.: Exploratory topic modeling with distributional semantics. In: Fromont, E., De Bie, T., van Leeuwen, M. (eds.) Advances in Intelligent Data Analysis XIV: 14th International Symposium, IDA 2015, Saint Etienne. France, October 22 -24, 2015. Proceedings. pp. 241–252. Springer International Publishing (2015)

[13] Scherer, M., von Landesberger, T., Schreck, T.: Topic modeling for search and exploration in multivariate research data repositories. In: Aalberg, T., Papatheodorou, C., Dobreva, M., Tsakonas, G., Farrugia, C.J. (eds.) Research and Advanced Technology for Digital Libraries: International Conference on Theory and Practice of Digital Libraries, TPDL 2013, Valletta, Malta, September 22-26, 2013. Proceedings. pp. 370–373. Springer Berlin Heidelberg, Berlin, Heidelberg (2013)

[14] Tan, Y., Ou, Z.: Topic-weak-correlated latent Dirichlet allocation. In: 7th International Symposium Chinese Spoken Language Processing (ISCSLP). pp. 224–228 (2010)

[15] Veas, E.E., di Sciascio, C.: Interactive topic analysis with visual analytics and recommender systems. In: 2nd Workshop on Cognitive Computing and Applications for Augmented Human Intelligence, CCAAHI2015, International Joint Conference on Artificial Intelligence, IJCAI, Buenos Aires, Argentina, July 2015. CEUR-WS.org, Aachen, Germany, Germany (2015)

[16] Vorontsov, K.V., Potapenko, A.A.: Tutorial on probabilistic topic modeling: Additive regularization for stochastic matrix factorization. In: AIST'2014, Analysis of Images, Social networks and Texts. vol. 436, pp. 29–46. Springer International Publishing Switzerland, Communications in Computer and Information Science (CCIS) (2014)

[17] Vorontsov, K.V., Potapenko, A.A.: Additive regularization of topic models. Machine Learning, Special Issue on Data Analysis and Intelligent Optimization with Applications 101(1), 303–323 (2015)

[18] Vorontsov, K.V., Potapenko, A.A., Plavin, A.V.: Additive regularization of topic models for topic selection and sparse factorization. In: et al., A.G. (ed.) The Third International Symposium On Learning And Data Sciences (SLDS 2015). April 20-22, 2015. Royal Holloway, University of London, UK. pp. 193–202. Springer International Publishing Switzerland 2015 (2015)

[19] Vorontsov, K., Frei, O., Apishev, M., Romov, P., Suvorova, M., Yanina, A.: Non-bayesian additive regularization for multimodal topic modeling of large collections. In: Proceedings of the 2015 Workshop on Topic Models: Post-Processing and Applications. pp. 29–37. ACM, New York, NY, USA (2015)

[20] Wei, X., Croft, W.B.: 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. SIGIR '06, ACM, New York, NY, USA (2006)

[21] White, R.W., Roth, R.A.: Exploratory Search: Beyond the Query-Response Paradigm. Synthesis Lectures on Information Concepts, Retrieval, and Services, Morgan and Claypool Publishers (2009)

[22] Yi, X., Allan, J.: A comparative study of utilizing topic models for information retrieval. In: Advances in Information Retrieval, Lecture Notes in Computer Science, vol. 5478, pp. 29–41. Springer Berlin Heidelberg (2009)