Automatic filtering of Russian scientific content using Machine Learning and Topic Modeling

Konstantin Vorontsov
CC RAS • Yandex • MIPT • HSE • MSU

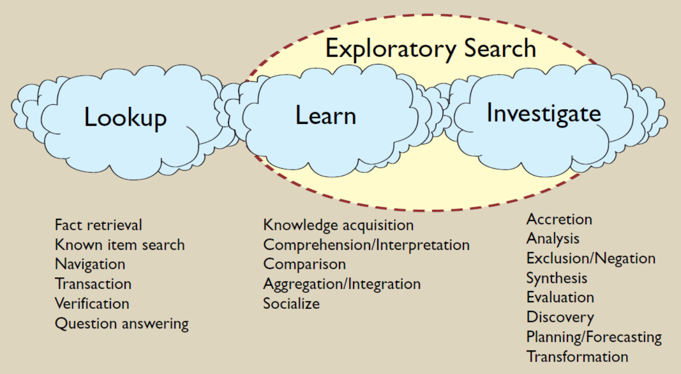
Sergey Voronov
MIPT • Skoltech • Yandex School of Data Analysis

• International Conference on Computational Linguistics •
Dialogue 2015 (May 27–30, Moscow)

Contents

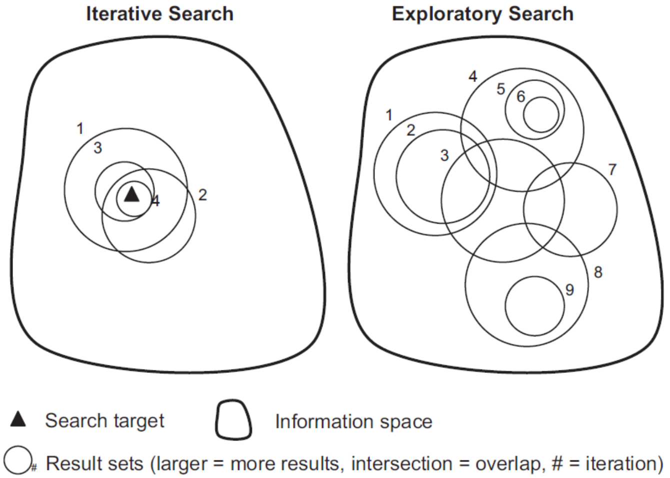
Exploratory Search for learning and investigation

- what if the user doesn't know which keywords to use?
- what if the user isn't looking for a single answer?



Gary Marchionini. Exploratory Search: from finding to understanding. Communications of the ACM. 2006, 49(4), p. 41–46.

Iterative “query-browse-refine” search vs Exploratory Search



R.W.White, R.A.Roth. Exploratory Search: beyond the Query-Response paradigm. San Rafael, CA: Morgan and Claypool, 2009.

Exploratory search scenario

Search query:

- a document of any length or even a set of documents

Search intents:

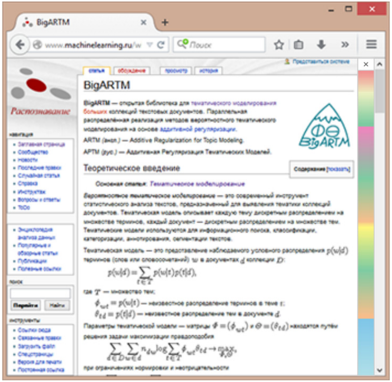
- what topics does it contain?
- what else is known on these topics?
- what is the structure of this domain area?
- what is most important, useful, popular, recent here?

Search scenario:

1. given a text (of any length) at hand (in any application)
2. identify topics and sub-topics it contains
3. show textual and graphical representations of these topics

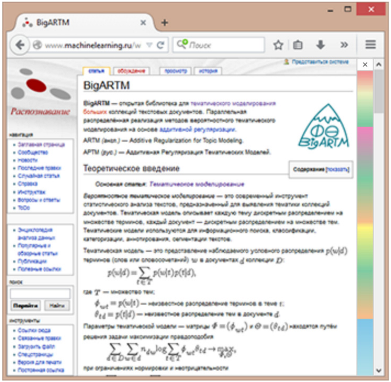
Exploratory search: the prototype of graphical user interface

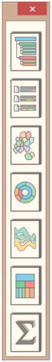
Color topic bar is a starting GUI element for exploratory search



Exploratory search: the prototype of graphical user interface

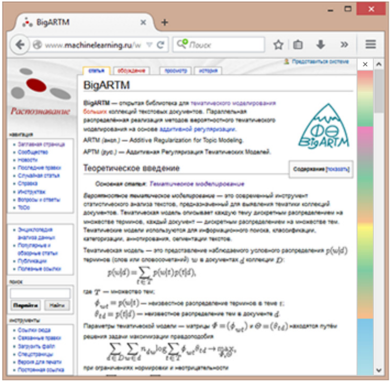
Click on the color topic bar is a topic query

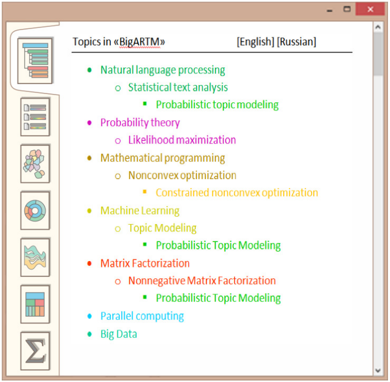




Exploratory search: the prototype of graphical user interface

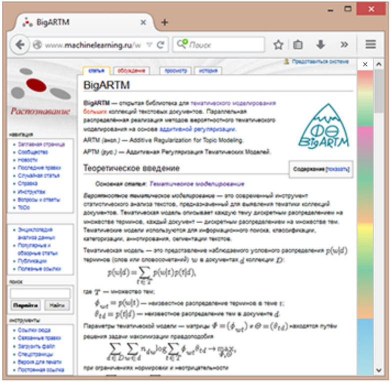
Topics of the query document

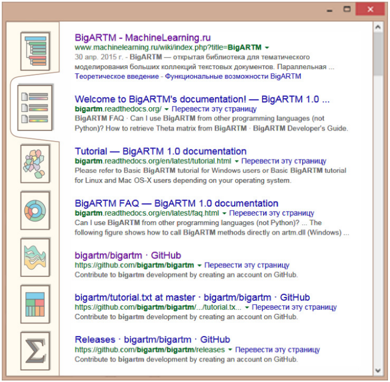




Exploratory search: the prototype of graphical user interface

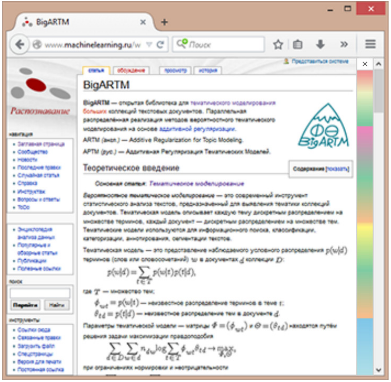
Documents and objects ranked by relevance

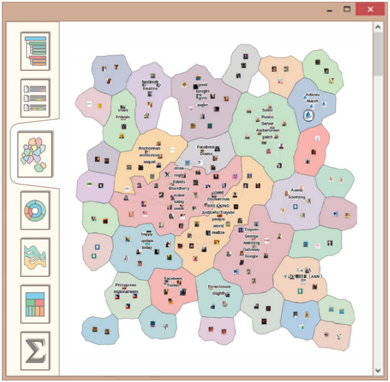




Exploratory search: the prototype of graphical user interface

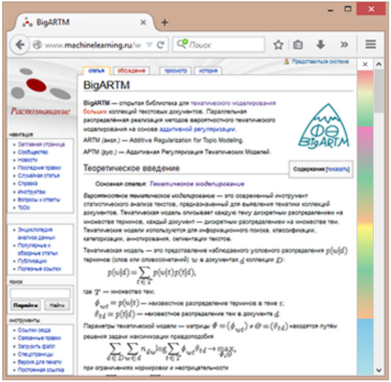
Topic roadmap: clustering of relevant documents

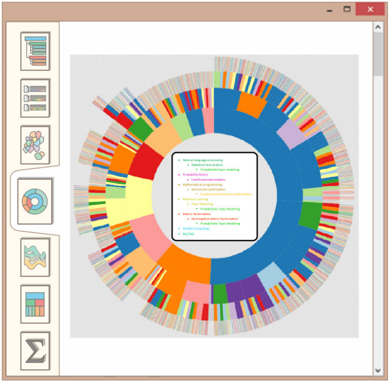




Exploratory search: the prototype of graphical user interface

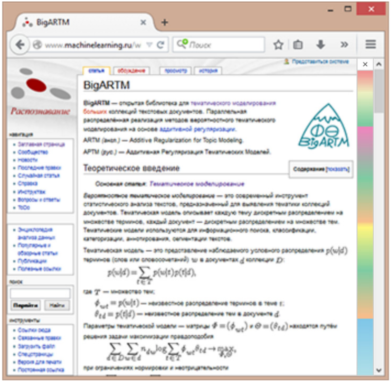
Topic hierarchy: topical structure of the domain area

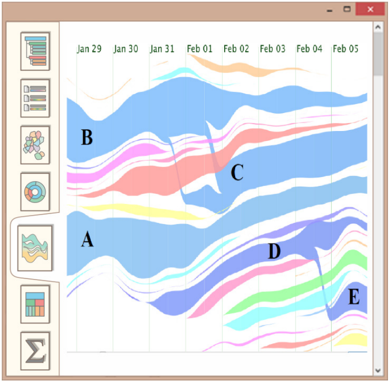




Exploratory search: the prototype of graphical user interface

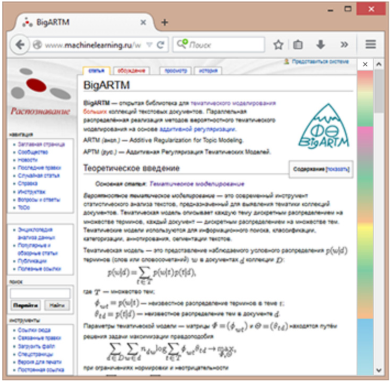
Topic river: evolution of the domain area

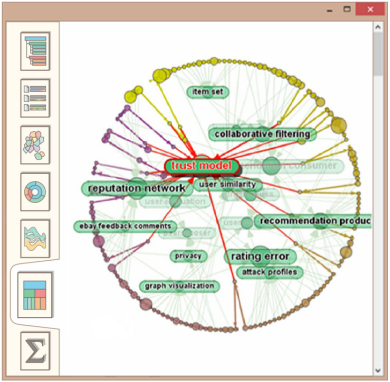




Exploratory search: the prototype of graphical user interface

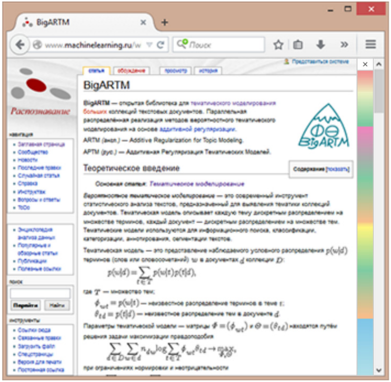
Topic bar: segmentation of the query document

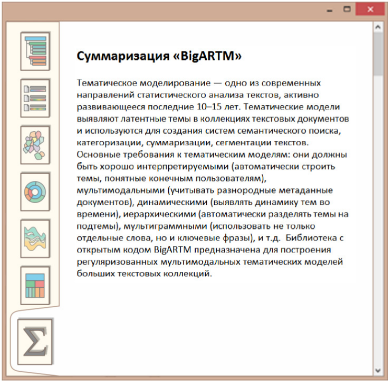




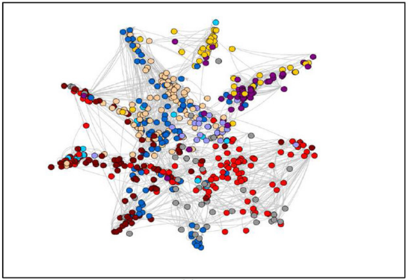
Exploratory search: the prototype of graphical user interface

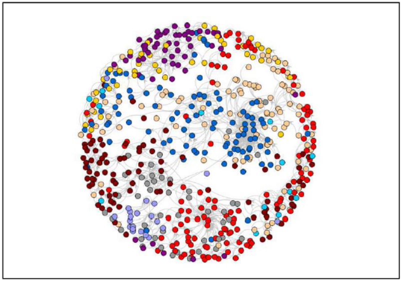
Summarization of the query document





Topic roadmap: clustering of relevant documents

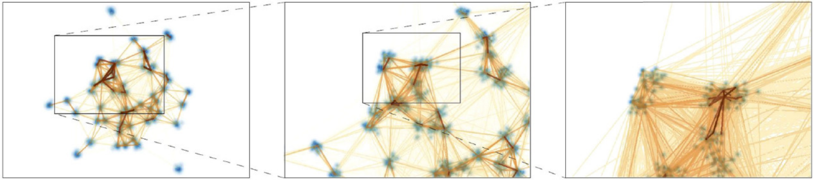




- Points represent documents
- Clusters represent groups of similar documents
- The most convenient shape of a cloud may be adjusted

Tuan M. V. Le, Hady W. Lauw Probabilistic Latent Document Network Embedding. IEEE International Conference ICDM. 2014.

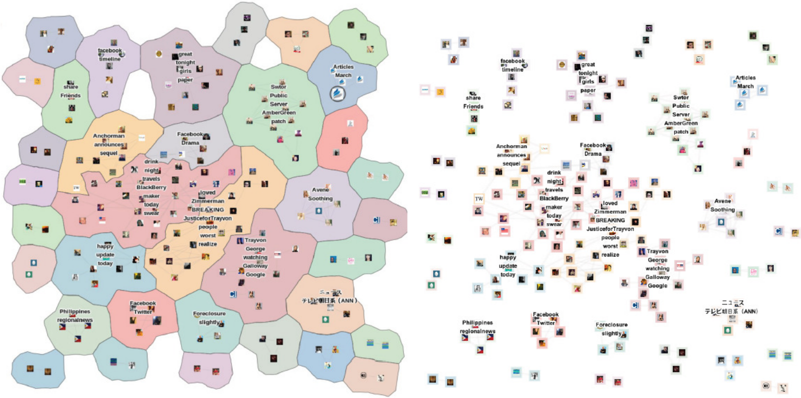
Topic roadmap: clustering of relevant documents



- Clusters
  of clusters
    of clusters
      of clusters ...

M.Zinsmaier, U.Brandes, O.Deussen, H.Strobelt. Interactive level-of-detail rendering of large graphs. IEEE Trans. Vis. Comput. Graph. 2012.

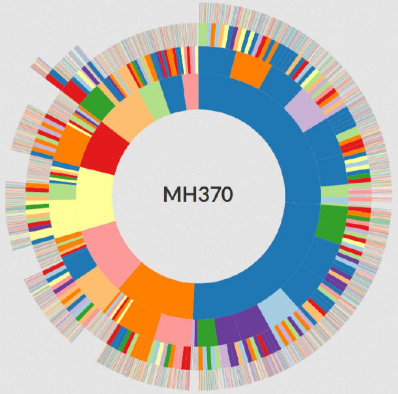
Topic roadmap: clustering of relevant documents



A map metaphor visualization (left) seems more appealing than a plain graph layout (right), and clusters seem easier to identify.

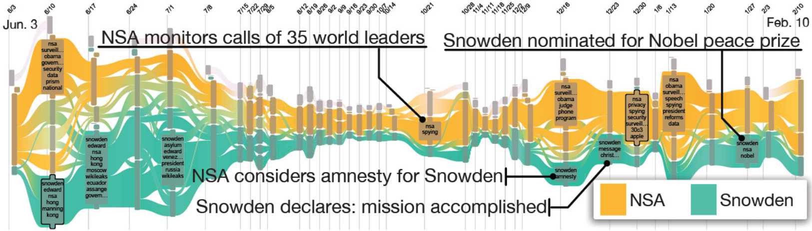
E.R.Gansner, Y.Hu, S.North. Visualizing Streaming Text Data with Dynamic Maps. 2012.

Topic hierarchy: topical structure of the domain area



Smith A., Hawes T., Myers M.. Hiérarchie: interactive visualization for hierarchical topic models. Workshop on Interactive Language Learning, Visualization, and Interfaces, ACL, 2014.

Topic river: evolution of the domain area

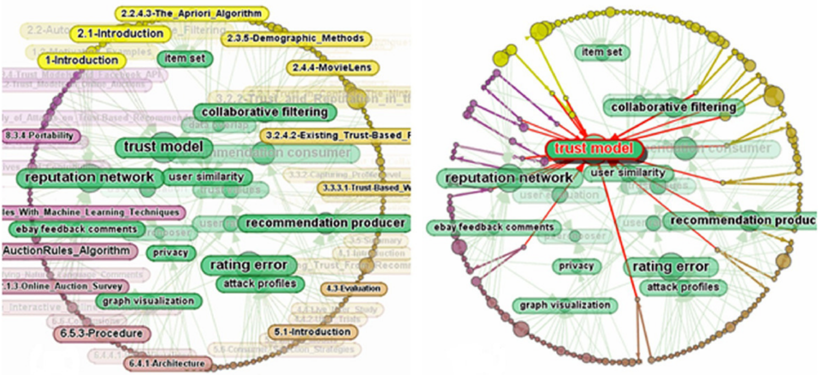


Evolving hierarchical topics in the Prism dataset (2013/06/03 – 2014/02/09).

- An expert chooses the cut of the tree hierarchy,
- marks events interactively,
- then generates a report.

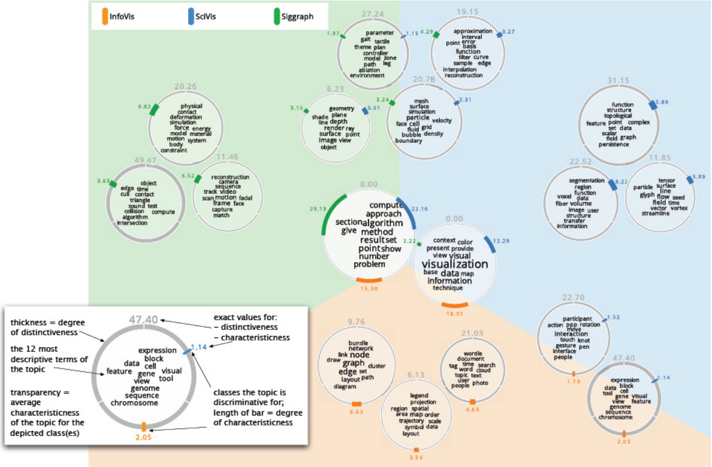
Weiwei Cui, Shixia Liu, Zhuofeng Wu, Hao Wei. How hierarchical topics evolve in large text corpora. IEEE Trans. Vis. Comput. Graph. 2014.

Topic bar: segmentation of the query document



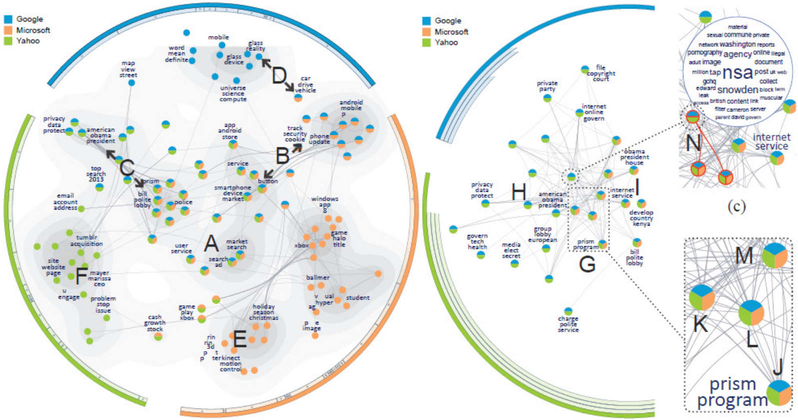
Gretarsson B., O'Donovan J., Bostandjiev S., Hollerer T., Asuncion A., Newman D., Smyth P. TopicNets: visual analysis of large text corpora with topic modeling. ACM Trans. on Intelligent Systems and Technology. 2012.

Topic sources: common topics and source-specific topics



Oelke D., Strobelt H., Rohrdantz C., Gurevych I., Deussen O. Comparative exploration of document collections: a visual analytics approach. EuroVis. 2014.

Topic sources: common topics and source-specific topics



Shixia Liu, Xiting Wang, Jianfei Chen, Jun Zhu, Baining Guo. TopicPanorama: a full picture of relevant topics. IEEE Symp. on Visual Analytics Science and Technology. 2014.

http://textvis.lnu.se

A visual survey of 170 text visualization techniques

The elements of Exploratory Search

1. Web crawling
2. Content filtering
3. Topic modeling
4. Building the inverted index
5. Ranking
6. Visualization

The elements of Exploratory Search

1. Web crawling . . . . ready-made solutions
2. Content filtering
3. Topic modeling
4. Building the inverted index . . . . ready-made solutions
5. Ranking . . . . ready-made solutions
6. Visualization . . . . ready-made solutions

The elements of Exploratory Search

1. Web crawling .... ready-made solutions
2. Content filtering .... **in this presentation**
3. Topic modeling .... **in this presentation**
4. Building the inverted index .... ready-made solutions
5. Ranking .... ready-made solutions
6. Visualization .... ready-made solutions

What is “topic”?

- *Topic* is a specific terminology of a particular domain area.
- *Topic* is a set of coherent terms (words or phrases) that often co-occur in documents.

More formally,

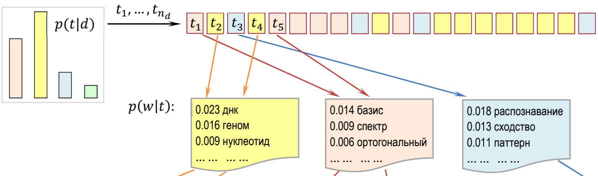
- *topic* is a probability distribution over terms: $p(w|t)$ is (unknown) frequency of word w in topic t.
- *document profile* is a probability distribution over *topics*: $p(t|d)$ is (unknown) frequency of topic t in document d.

When writing term w in document d author thinks of topic t.
Topic model tries to uncover latent topics in a text collection.

Probabilistic Topic Model (PTM)

Topic model explains terms w in documents d by topics t:

$$p(w|d) = \sum_t p(w|t)p(t|d)$$

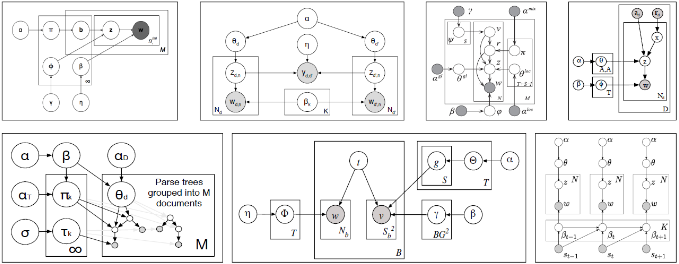


$w_1, \dots, w_{n_d}$:

Разработан спектрально-аналитический подход к выявлению размытых протяженных повторов в геномных последовательностях. Метод основан на разномасштабном оценивании сходства нуклеотидных последовательностей в пространстве коэффициентов разложения фрагментов кривых GC- и GA-содержания по классическим ортогональным базисам. Найдены условия оптимальной аппроксимации, обеспечивающие автоматическое распознавание повторов различных видов (прямых и инвертированных, а также тандемных) на спектральной матрице сходства. Метод одинаково хорошо работает на разных масштабах данных. Он позволяет выявлять следы сегментных дупликаций и мегасателлитные участки в геноме, районы синтении при сравнении пары геномов. Его можно использовать для детального изучения фрагментов хромосом (поиска размытых участков с умеренной длиной повторяющегося паттерна).

Probabilistic topic modeling: milestones and mainstream

1. PLSA — Probabilistic Latent Semantic Analysis (1999)
2. LDA — Latent Dirichlet Allocation (2003)
3. 100s of PTMs based on Graphical Models & Bayesian Inference



David Blei. Probabilistic topic models // Communications of the ACM, 2012. Vol. 55. No. 4. Pp. 77–84.

Topic model for exploratory search should be...

1. **Interpretable**: each topic should be well interpretable by humans and labeled

Topic model for exploratory search should be...

1. **Interpretable**: each topic should be well interpretable by humans and labeled
2. **Multigram**: keyphrases should be extracted automatically

Topic model for exploratory search should be...

1. **Interpretable**: each topic should be well interpretable by humans and labeled
2. **Multigram**: keyphrases should be extracted automatically
3. **Multilingual**: cross-language and multi-language search should be supported

Topic model for exploratory search should be...

1. **Interpretable**: each topic should be well interpretable by humans and labeled
2. **Multigram**: keyphrases should be extracted automatically
3. **Multilingual**: cross-language and multi-language search should be supported
4. **Multimodal**: authors, categories, sources, links, tags, named entities, users, etc. should be involved in the model

Topic model for exploratory search should be...

1. **Interpretable**: each topic should be well interpretable by humans and labeled
2. **Multigram**: keyphrases should be extracted automatically
3. **Multilingual**: cross-language and multi-language search should be supported
4. **Multimodal**: authors, categories, sources, links, tags, named entities, users, etc. should be involved in the model
5. **Temporal**: topic dynamics over time should be identified

Topic model for exploratory search should be...

1. **Interpretable**: each topic should be well interpretable by humans and labeled
2. **Multigram**: keyphrases should be extracted automatically
3. **Multilingual**: cross-language and multi-language search should be supported
4. **Multimodal**: authors, categories, sources, links, tags, named entities, users, etc. should be involved in the model
5. **Temporal**: topic dynamics over time should be identified
6. **Hierarchical**: granularity of topics should be user-adjustable

Topic model for exploratory search should be...

1. **Interpretable**: each topic should be well interpretable by humans and labeled
2. **Multigram**: keyphrases should be extracted automatically
3. **Multilingual**: cross-language and multi-language search should be supported
4. **Multimodal**: authors, categories, sources, links, tags, named entities, users, etc. should be involved in the model
5. **Temporal**: topic dynamics over time should be identified
6. **Hierarchical**: granularity of topics should be user-adjustable
7. **Segmented**: the topical text segmentation should be supported beyond the bag-of-words (BoW) model

Topic model for exploratory search should be...

1. **Interpretable**: each topic should be well interpretable by humans and labeled
2. **Multigram**: keyphrases should be extracted automatically
3. **Multilingual**: cross-language and multi-language search should be supported
4. **Multimodal**: authors, categories, sources, links, tags, named entities, users, etc. should be involved in the model
5. **Temporal**: topic dynamics over time should be identified
6. **Hierarchical**: granularity of topics should be user-adjustable
7. **Segmented**: the topical text segmentation should be supported beyond the bag-of-words (BoW) model
8. **Semi-supervised**: the corrections from experts should be used to improve the model

What prevents usage of topic modeling for exploratory search?

1. Lack of techniques for combining topic models

How are we going to solve these problems:

1. ARTM — Additive Regularization for Topic Modeling

What prevents usage of topic modeling for exploratory search?

1. Lack of techniques for combining topic models
2. Lack of interpretability

How are we going to solve these problems:

1. ARTM — Additive Regularization for Topic Modeling
2. - Automatic multigram term extraction
   - Using external linguistic resources (thesaurus, ontologies)
   - Automatic revealing of topic lexical kernels (via sparsity, diversity and coherence maximization)

What prevents usage of topic modeling for exploratory search?

1. Lack of techniques for combining topic models
2. Lack of interpretability
3. Lack of linguistic validity

How are we going to solve these problems:

1. ARTM — Additive Regularization for Topic Modeling
2.
   - Automatic multigram term extraction
   - Using external linguistic resources (thesaurus, ontologies)
   - Automatic revealing of topic lexical kernels (via sparsity, diversity and coherence maximization)
3. Linguistic regularization of topic models (sentence TM, syntactic TM, segmentation TM, etc.)

ARTM: Additive Regularization for Topic Modeling

Given: W is a set (vocabulary) of terms
D is a set (collection) of documents $d = \{w_1 \dots w_{n_d}\}$
n_{dw} = how many times term w appears in document d

Find: parameters $\phi_{wt} = p(w|t)$, $\theta_{td} = p(t|d)$ of the topic model

The problem is to maximize the *regularized* log-likelihood:

$$\sum_{d,w} n_{dw} \ln \sum_t \phi_{wt}\theta_{td} + R(\Phi, \Theta) \ \to\ \max_{\Phi,\Theta},$$

$$\phi_{wt} \geqslant 0, \quad \sum_{w\in W} \phi_{wt} = 1; \qquad \theta_{td} \geqslant 0, \quad \sum_{t\in T} \theta_{td} = 1.$$

Vorontsov K. V., Potapenko A. A. Tutorial on probabilistic topic modeling: additive regularization for stochastic matrix factorization // AIST'2014.

Solution: the regularized EM algorithm

Input: collection D: each $d = \{w_1 \dots w_{n_d}\}$ also as BoW $\|n_{dw}\|$;
Output: matrices $\phi_{wt} = p(w|t)$, $\theta_{td} = p(t|d)$;

initialize ϕ_{wt}, θ_{td};
repeat
  estimate topic distribution for each term w in each document d:
  $p(t|d,w) = \mathrm{norm}_t\bigl(\phi_{wt}\theta_{td}\bigr)$;
  count the frequency of each term w in each topic t:
  $n_{wt} = \sum_d n_{dw} p(t|d,w)$;
  count the frequency of each topic t in each document d:
  $n_{td} = \sum_w n_{dw} p(t|d,w)$;
  apply regularization and normalize conditional probabilities:
  $\phi_{wt} = \mathrm{norm}_w\Bigl(n_{wt} + \phi_{wt}\frac{\partial R}{\partial \phi_{wt}}\Bigr)$;
  $\theta_{td} = \mathrm{norm}_t\Bigl(n_{td} + \theta_{td}\frac{\partial R}{\partial \theta_{td}}\Bigr)$;
until convergence;

ARTM: available regularizers

- topic smoothing (equivalent to LDA)
- topic sparsing
- topic decorrelation
- topic selection via entropy sparsing
- topic coherence maximization
- supervised learning for classification and regression
- semi-supervised learning
- using documents citation and links
- modeling temporal topic dynamics
- using vocabularies in multilingual topic models
- and many others

Vorontsov K. V., Potapenko A. A. Additive Regularization of Topic Models // Machine Learning. Special Issue "Data Analysis and Intelligent Optimization with Applications". Springer, 2014.

Summary of ARTM approach

EM-algorithm is computationally effecient:

- It has linear time complexity $O(n \cdot |T| \cdot \text{nIter})$
- Its online version passes only once through a big collection
- Parallelism is possible for both multi-core CPUs and clusters

ARTM reduces barriers to entry into PTM research field:

- PLSA, LDA, and 100s of PTMs are covered by ARTM
- Combining multiple regularizers is easy
- No complicated Bayesian inference and graphical models
- Fast parallel online implementation BigARTM (bigartm.org)

Next step: making topic models more linguistic

- Syntactic topic models
- Sentence and discourse topic models
- Text segmentation topic models
- Automatic multigram term extraction
- ARTM: post-processing of topic–term matrix $p(t|d,w)$

J.Boyd-Graber. Linguistic extensions of topic models. PhD thesis. 2010.
N.Aletras. Interpreting document collections with topic models. PhD thesis. 2014.
M.Yang, T.Cui, W.Tu. Ordering-sensitive and semantic-aware topic modeling. 2015.
M.Riedl, C.Biemann. How text segmentation algorithms gain from topic models. 2012.
S.Remus, C.Biemann Three knowledge-free methods for automatic lexical chain extraction. 2013.
A.Lazaridou, I.Titov, C.Sporleder. A Bayesian model for joint unsupervised induction of sentiment, aspect and discourse representations. 2013.

BigARTM project

BigARTM features:

- Parallel + Online + Multimodal + Regularized topic modeling
- Out-of-core processing of Big Data
- Built-in library of regularizers and quality measures

BigARTM project

BigARTM features:

- Parallel + Online + Multimodal + Regularized topic modeling
- Out-of-core processing of Big Data
- Built-in library of regularizers and quality measures

BigARTM community:

- Open-source https://github.com/bigartm
  (discussion group, issue tracker, pull requests)
- Documentation http://bigartm.org

BigARTM project

BigARTM features:

- Parallel + Online + Multimodal + Regularized topic modeling
- Out-of-core processing of Big Data
- Built-in library of regularizers and quality measures

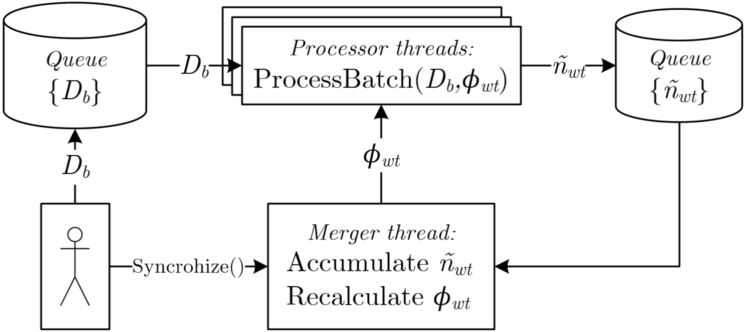
BigARTM community:

- Open-source https://github.com/bigartm
  (discussion group, issue tracker, pull requests)
- Documentation http://bigartm.org

BigARTM license and programming environment:

- Freely available for commercial usage (BSD 3-Clause license)
- Cross-platform — Windows, Linux, Mac OS X (32 bit, 64 bit)
- Programming APIs: command-line, C++, and Python

The BigARTM project: parallel architecture



- Concurrent processing of batches
- Simple single-threaded code for *ProcessBatch*
- User controls when to update the model in online algorithm
- Deterministic (reproducible) results from run to run

Experiment 1. BigARTM vs Gensim vs Vowpal Wabbit

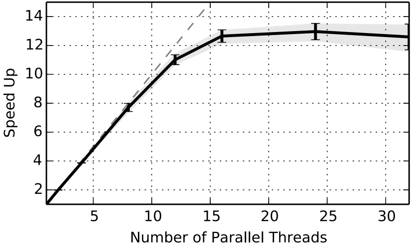
- 3.7M articles from Wikipedia, 100K unique words

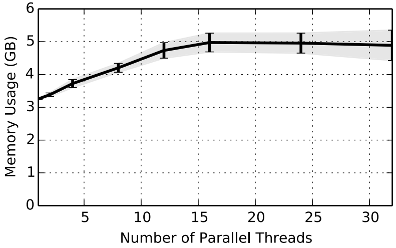
	procs	train	inference	perplexity
BigARTM	1	35 min	72 sec	4000
Gensim.LdaModel	1	369 min	395 sec	4161
VowpalWabbit.LDA	1	73 min	120 sec	4108
BigARTM	4	9 min	20 sec	4061
Gensim.LdaMulticore	4	60 min	222 sec	4111
BigARTM	8	4.5 min	14 sec	4304
Gensim.LdaMulticore	8	57 min	224 sec	4455

- *procs* = number of parallel threads
- *inference* = time to infer θ_d for 100K held-out documents
- *perplexity* is calculated on held-out documents.

Experiment 1. Running BigARTM in parallel

- 3.7M articles from Wikipedia, 100K unique words



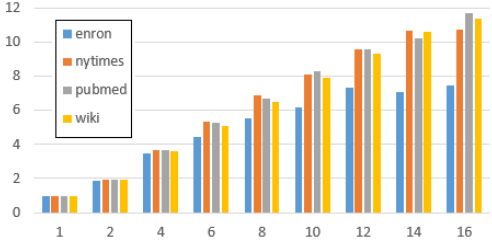


- Amazon EC2 c3.8xlarge (16 physical cores + hyperthreading)
- No extra memory cost for adding more threads

Experiment 2. Running BigARTM on large collections

collection	$\|W\|$, 10^3	$\|D\|$, 10^6	n, 10^6	size, GB
enron	28	0.04	6.4	0.07
nytimes	103	0.3	100	0.13
pubmed	141	8.2	738	1.0
wiki	100	3.7	1009	1.2

speedup

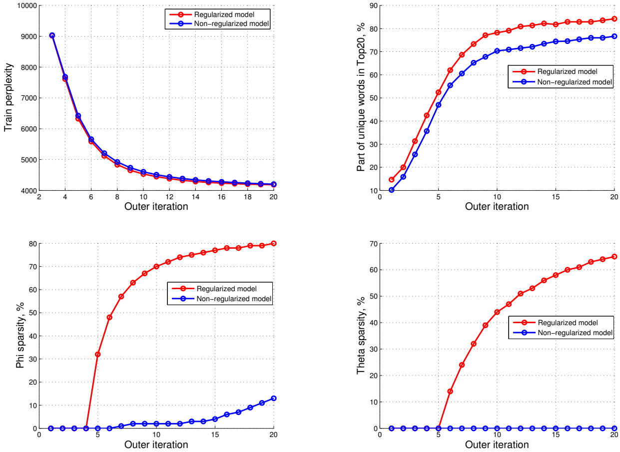


processors

Amazon EC2 cc2.8xlarge instance:
16 cores + hyperthreading, Intel® Xeon® CPU E5-2670 2.6GHz.

Experiment 3. Additive regularization

ARTM combines regularizers to improve multiple criteria (sparsity, number of unique words) without a loss of the perplexity.



Experiment 4. Interpretability of Multilingual ARTM

We consider languages as modalities in Multimodal ARTM.

Collection of 216 175 Russian–English Wikipedia articles pairs.
Top 10 words with $p(w|t)$ probabilities (in %):

Topic 68				Topic 79			
research	4.56	институт	6.03	goals	4.48	матч	6.02
technology	3.14	университет	3.35	league	3.99	игрок	5.56
engineering	2.63	программа	3.17	club	3.76	сборная	4.51
institute	2.37	учебный	2.75	season	3.49	фк	3.25
science	1.97	технический	2.70	scored	2.72	против	3.20
program	1.60	технология	2.30	cup	2.57	клуб	3.14
education	1.44	научный	1.76	goal	2.48	футболист	2.67
campus	1.43	исследование	1.67	apps	1.74	гол	2.65
management	1.38	наука	1.64	debut	1.69	забивать	2.53
programs	1.36	образование	1.47	match	1.67	команда	2.14

Experiment 4. Interpretability of Multilingual ARTM

Collection of 216 175 Russian–English Wikipedia articles pairs.
Top 10 words with $p(w|t)$ probabilities (in %):

Topic 88				Topic 251			
opera	7.36	опера	7.82	windows	8.00	windows	6.05
conductor	1.69	оперный	3.13	microsoft	4.03	microsoft	3.76
orchestra	1.14	дирижер	2.82	server	2.93	версия	1.86
wagner	0.97	певец	1.65	software	1.38	приложение	1.86
soprano	0.78	певица	1.51	user	1.03	сервер	1.63
performance	0.78	театр	1.14	security	0.92	server	1.54
mozart	0.74	партия	1.05	mitchell	0.82	программный	1.08
sang	0.70	сопрано	0.97	oracle	0.82	пользователь	1.04
singing	0.69	вагнер	0.90	enterprise	0.78	обеспечение	1.02
operas	0.68	оркестр	0.82	users	0.78	система	0.96

All $|T| = 400$ topics were reviewed by an independent assessor, and he successfully interpreted 396 topics.

Experiment 5. Interpretability of Multigram ARTM

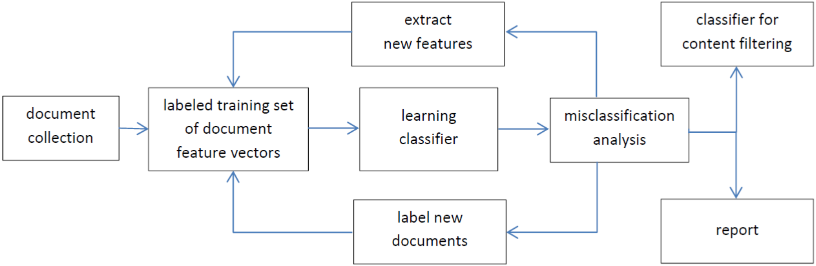
We consider *n*-grams as modalities in Multimodal ARTM.

Collection: 1000 articles from Russian conference www.mmro.ru

распознавание образов в биоинформатике		теория вычислительной сложности	
unigrams	bigrams	unigrams	bigrams
объект	задача распознавания	задача	разделять множества
задача	множество мотивов	множество	конечное множество
множество	система масок	подмножество	условие задачи
мотив	вторичная структура	условие	задача о покрытии
разрешимость	структура белка	класс	покрытие множества
выборка	распознавание вторичной	решение	сильный смысл
маска	состояние объекта	конечный	разделяющий комитет
распознавание	обучающая выборка	число	минимальный аффинный
информативность	оценка информативности	аффинный	аффинный комитет
состояние	множество объектов	случай	аффинный разделяющий
закономерность	разрешимость задачи	покрытие	общее положение
система	критерий разрешимости	общий	множество точек
структура	информативность мотива	пространство	случай задачи
значение	первичная структура	схема	общий случай
регулярность	тупиковое множество	комитет	задача MASC

Learning the Content Filter

1. the principles of filtering are communicated to the experts
2. experts label the training set as «good» and «bad» documents
3. the learned classifier is used for *uncertainty sampling*



The problem of unbalanced classes: 1 : 50 in our collection

Topic modeling for scientific genre classification

Collection:
850K documents from 2000 sites of Russian universities,
no more than 2% of them are «good» i.e. scientific.

Labeled: 3K documents, 40% of them are «good».

Expert instructions:

- «good» document is a primary source of scientific information that carries scientific knowledge
- «bad» documents: commercial and organizational content, home pages, courses, annotations, bibliographies, etc.

Topic model for semi-supervised classification
learns positive and negative words for text genre classification from a small labeled training set.

Examples of topics

← less scientific | more scientific →

0.000	0.099	0.203	0.755	1.000
образование	который	страна	прямая	процесс
студент	ряд	Россия	быть	результат
социальный	оценка	производство	точка	модель
учебный	человек	экономика	значение	среда
современный	параметр	который	множество	зависимость
университет	источник	год	движение	различный
российский	система	орган	состояние	структура
научный	помощь	государство	система	позволять
формирование	связь	проблема	время	являться
вуз	этот	развитие	рис	поверхность
конференция	комплекс	период	коэффициент	расчет
организация	наличие	федеральный	тогда	технический
проект	изменение	закон	исследование	обработка
история	труд	хозяйство	свойство	качество
место	мир	такой	граница	данный
вуз	знание	власть	вектор	моделирование
кафедра	высокий	стоимость	уровень	сигнал
личность	величина	весь	коэффициент	следующий
субъект	число	условие	вероятность	основа

Content filtering results

F1, Recall, Precision (in %) for balanced and unbalanced data:

features	F1	Recall	Prec	F1	Recall	Prec
A1	76.57	62.08	99.90	77.40	71.48	84.39
A1 T	93.27	91.54	95.07	77.38	72.35	83.13
A1 A2	93.18	90.63	95.87	81.86	83.33	80.06
A1 A2 T	92.62	90.13	95.24	81.95	80.21	83.76
A1 A2 T8	95.12	95.52	94.72	82.24	78.54	86.31
A1 A2 B	96.24	95.92	96.57	90.37	91.00	89.75
A1 A2 B T	96.50	96.22	96.78	90.51	93.21	87.95
A1 A2 B T8	97.33	97.47	97.20	90.85	93.42	88.41

Groups of features:

- **A1** Greek letters, math symbols, document length
- **A2** digits, short text indicator, grant phrase words
- **B** bibliography lines, references
- **T** 25 topics
- **T8** 8 topics selected manually from several runs

Conclusions

- Topic modeling & content filtering
  are key technologies for exploratory search
- ARTM (Additive Regularization for Topic Modeling)
  is a general framework, which makes topic models
  easy to design, to infer, to explain, and to combine.
- BigARTM is an open source project ready for parallel online multimodal topic modeling of large text collections.

http://bigartm.org

Join BigARTM community!