BigARTM: Open Source Library for Topic Modeling of Large Text Collections

Konstantin Vorontsov, Oleksandr Frei, Murat Apishev,
Peter Romov, Anastasia Yanina, Marina Suvorova

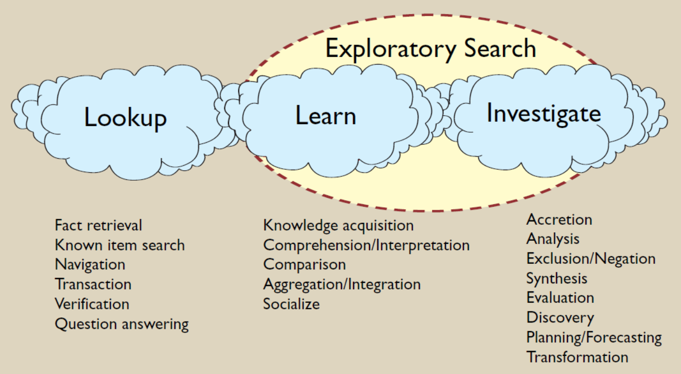
(CC RAS, MIPT, MSU, Yandex • Moscow, Russia)

XVII International Conference DAMDID/RCDL'2015
Data Analytics and Management in Data Intensive Domains
October 13–16, 2015 • Obninsk, Russia

Contents

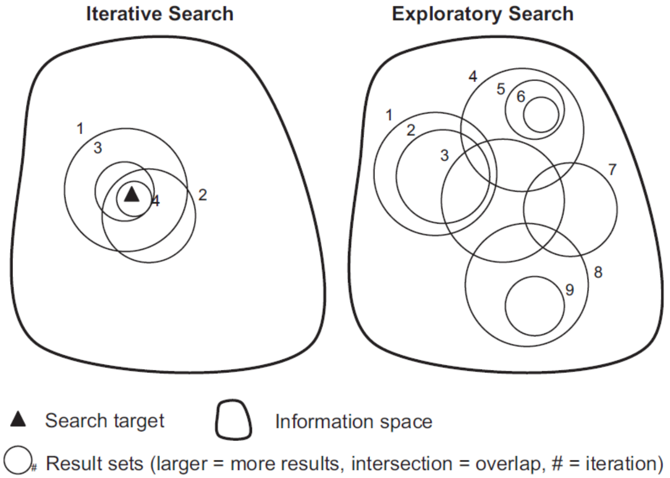
Exploratory Search for learning, knowledge acquisition and discovery

- 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:

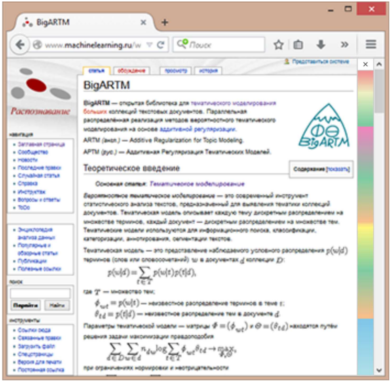
- what topics does it contain?
- what else is known on these topics?
- what is the structure of this domain area?
- what is most important, useful, popular, recent here?

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

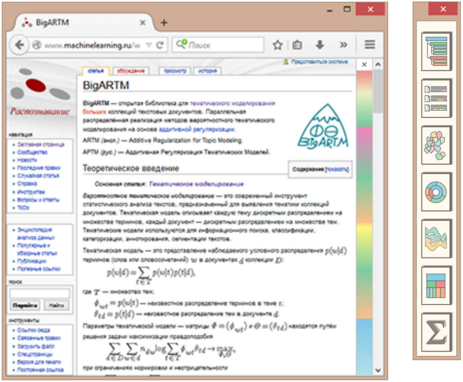
Exploratory search: the prototype of graphical user interface

Color topic bar is a starting GUI element for exploratory search



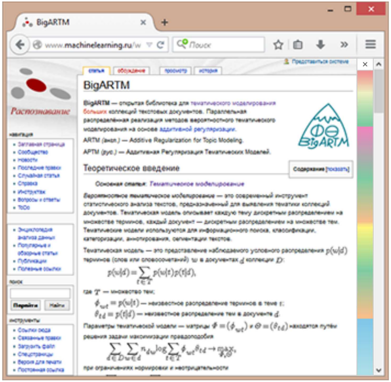
Exploratory search: the prototype of graphical user interface

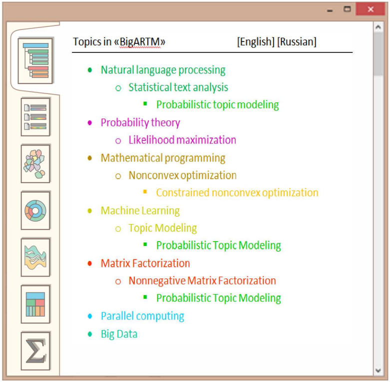
Click on the color topic bar is a topic query



Exploratory search: the prototype of graphical user interface

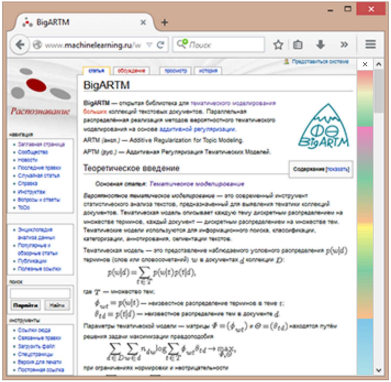
Topics of the query document

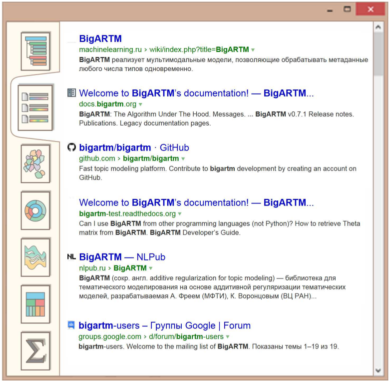




Exploratory search: the prototype of graphical user interface

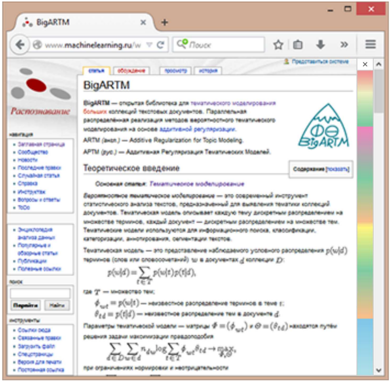
Similar documents and objects ranked by relevance

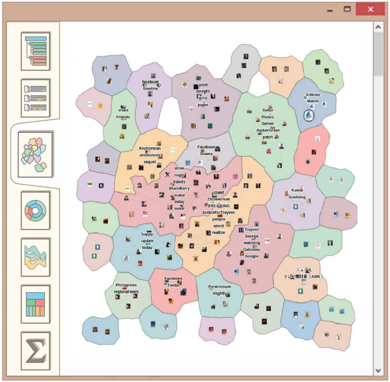




Exploratory search: the prototype of graphical user interface

Topic roadmap: clustering of relevant documents

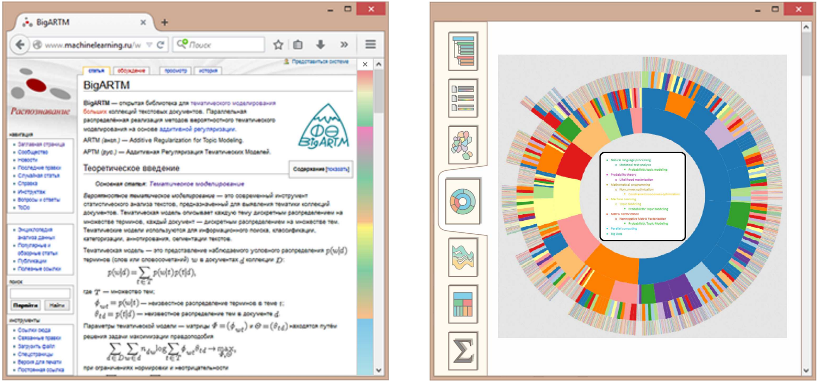




E.R.Gansner, Y.Hu, S.North. Visualizing Streaming Text Data with Dynamic Maps. 2012.

Exploratory search: the prototype of graphical user interface

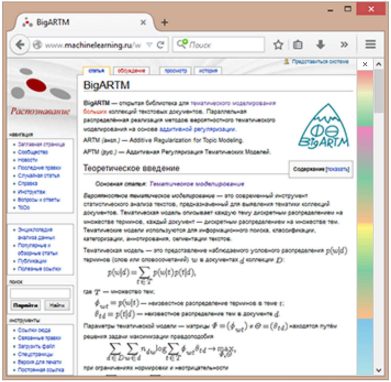
Topic hierarchy: topical structure of the domain area

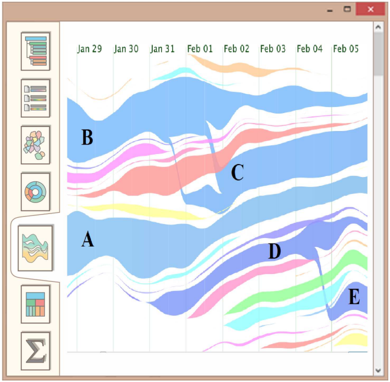


Smith A., Hawes T., Myers M.. Hiérarchie: interactive visualization for hierarchical topic models. Workshop on Interactive Language Learning, Visualization, and Interfaces, ACL, 2014.

Exploratory search: the prototype of graphical user interface

Topic river: evolution of the domain area

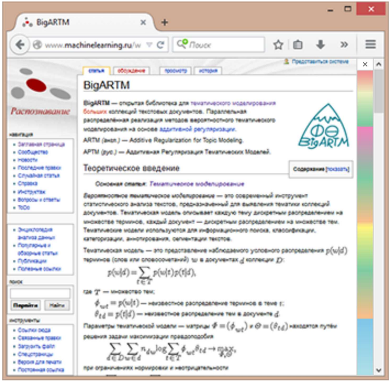


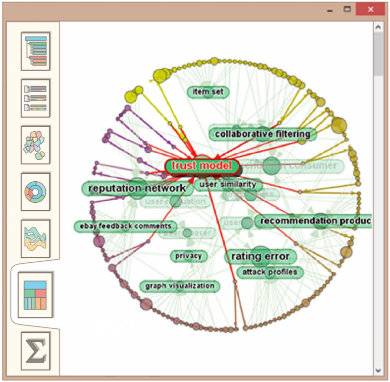


Weiwei Cui, Shixia Liu, Zhuofeng Wu, Hao Wei. How hierarchical topics evolve in large text corpora. IEEE Trans. Vis. Comput. Graph. 2014.

Exploratory search: the prototype of graphical user interface

Topic bar: segmentation of the query document

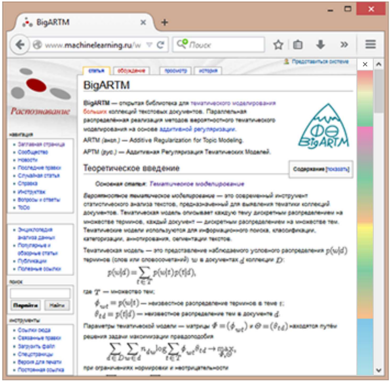


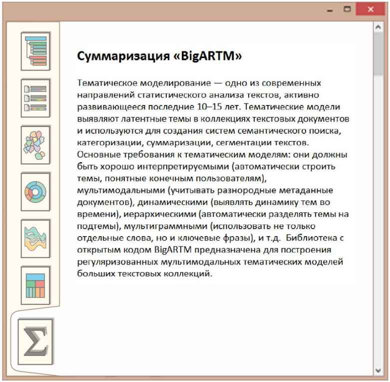


Gretarsson B., O'Donovan J., Bostandjiev S., Hollerer T., Asuncion A., Newman D., Smyth P. TopicNets: visual analysis of large text corpora with topic modeling. ACM Trans. on Intelligent Systems and Technology. 2012.

Exploratory search: the prototype of graphical user interface

Summarization of the query document





http://textvis.lnu.se

A visual survey of 220 text visualization techniques

The elements of Exploratory Search technology

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

Topic Model used for Exploratory Search must be...

1. **Interpretable**: each topic should be well interpretable by humans and labeled automatically
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 model
8. **Semi-supervised**: labeling should be used to improve the model
9. **Online, parallel, distributed**: big data should be processed

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,

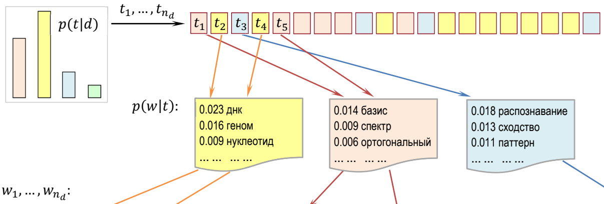
- *topic* is a probability distribution over terms:
  $p(w|t)$ is (unknown) frequency of word w in topic t.
- *document profile* is a probability distribution over *topics*:
  $p(t|d)$ is (unknown) frequency of topic t in document d.

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) generating a text collection

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

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



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).

Inverse problem: text collection $\to$ PTM

Given: D is a set (collection) of documents
W is a set (vocabulary) of terms
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

$$p(w|d) = \sum_t \phi_{wt}\theta_{td}.$$

under nonnegativity and normalization constraints

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

The ill-posed problem of matrix factorization:

$$\Phi\Theta = (\Phi S)(S^{-1}\Theta) = \Phi'\Theta'$$

for all S such that Φ', Θ' are stochastic.

PLSA — Probabilistic Latent Semantic Analysis [Hofmann, 1999]

Constrained maximization of the log-likelihood:

$$\mathscr{L}(\Phi, \Theta) = \sum_{d,w} n_{dw} \ln \sum_t \phi_{wt}\theta_{td} \;\to\; \max_{\Phi,\Theta}$$

EM-algorithm is a simple iteration method for the nonlinear system

$$\begin{array}{rl} \text{E-step:} & \left\{ \begin{array}{l} p_{tdw} \equiv p(t|d,w) = \operatorname*{norm}_{t\in T}\left(\phi_{wt}\theta_{td}\right) \\ \phi_{wt} = \operatorname*{norm}_{w\in W}\left(\sum\limits_{d\in D} n_{dw} p_{tdw}\right) \\ \theta_{td} = \operatorname*{norm}_{t\in T}\left(\sum\limits_{w\in d} n_{dw} p_{tdw}\right) \end{array} \right. \\ \text{M-step:} & \end{array}$$

where $\operatorname*{norm}_{t\in T} x_t = \frac{\max\{x_t,0\}}{\sum\limits_{s\in T}\max\{x_s,0\}}$ is vector normalization.

LDA — Latent Dirichlet Allocation [Blei, Ng, Jordan, 2003]

Maximum a posteriori (MAP) with Dirichlet prior:

$$\underbrace{\sum_{d,w} n_{dw} \ln \sum_{t} \phi_{wt}\theta_{td}}_{\text{log-likelihood } \mathscr{L}(\Phi,\Theta)} + \underbrace{\sum_{t,w} \beta_w \ln \phi_{wt} + \sum_{d,t} \alpha_t \ln \theta_{td}}_{\text{regularization criterion } R(\Phi,\Theta)} \to \max_{\Phi,\Theta}$$

EM-algorithm is a simple iteration method for the system

$$\begin{array}{rl} \text{E-step:} & \left\{ \begin{array}{l} p_{tdw} = \operatorname*{norm}_{t\in T}\big(\phi_{wt}\theta_{td}\big) \\ \phi_{wt} = \operatorname*{norm}_{w\in W}\Big(\sum\limits_{d\in D} n_{dw}p_{tdw} + \beta_w\Big) \\ \theta_{td} = \operatorname*{norm}_{t\in T}\Big(\sum\limits_{w\in d} n_{dw}p_{tdw} + \alpha_t\Big) \end{array} \right. \\ \text{M-step:} & \end{array}$$

ARTM — Additive Regularization of Topic Model [Vorontsov, 2014]

Maximum log-likelihood with additive regularization criterion R:

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

EM-algorithm is a simple iteration method for the system

$$\begin{array}{rl} \text{E-step:} & \left\{ \begin{array}{l} p_{tdw} = \operatorname*{norm}_{t\in T}\big(\phi_{wt}\theta_{td}\big) \\[2ex] \phi_{wt} = \operatorname*{norm}_{w\in W}\Big(\sum_{d\in D} n_{dw}p_{tdw} + \phi_{wt}\frac{\partial R}{\partial \phi_{wt}}\Big) \\[2ex] \theta_{td} = \operatorname*{norm}_{t\in T}\Big(\sum_{w\in d} n_{dw}p_{tdw} + \theta_{td}\frac{\partial R}{\partial \theta_{td}}\Big) \end{array} \right. \\ \text{M-step:} & \end{array}$$

Many Bayesian PTMs can be reinterpreted as regularizers in ARTM

- smoothing (LDA) for background and stop-words topics
- sparsing (anti-LDA) for domain-specific topics
- topic decorrelation
- topic coherence maximization
- supervised learning for classification and regression
- semi-supervised learning
- using document citations and links
- determining number of topics via entropy sparsing
- modeling topical hierarchies
- modeling temporal topic dynamics
- using vocabularies in multilingual topic models
- etc.

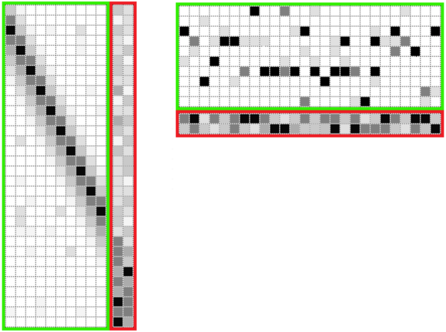
Vorontsov K. V., Potapenko A. A. Additive Regularization of Topic Models // Machine Learning. Volume 101, Issue 1 (2015), Pp. 303-323.

Assumptions: what topics would be well-interpretable?

Specific topics S contain domain-specific terms
$p(w|t)$ are sparse and different (weakly correlated)

Background topics B contain common lexis words
$p(w|t)$ are not sparse

ϕ_{wt} terms$\times$topics $\qquad$ θ_{td} topics$\times$documents



Smoothing regularization (rethinking LDA)

The non-sparsity assumption for background topics $t \in B$:
ϕ_{wt} are similar to a given distribution β_w;
θ_{td} are similar to a given distribution α_t.

Minimize the sum of KL-divergences $\mathrm{KL}(\beta\|\phi_t)$ and $\mathrm{KL}(\alpha\|\theta_d)$:

$$R(\Phi,\Theta) = \beta_0 \sum_{t\in B}\sum_{w\in W} \beta_w \ln \phi_{wt} + \alpha_0 \sum_{d\in D}\sum_{t\in B} \alpha_t \ln \theta_{td} \to \max .$$

The regularized M-step applied for all $t \in B$ coincides with LDA:

$$\phi_{wt} \propto n_{wt} + \beta_0\beta_w, \qquad \theta_{td} \propto n_{td} + \alpha_0\alpha_t,$$

which is new non-Bayesian interpretation of LDA [Blei 2003].

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

Sparsing regularizer (further rethinking LDA)

The sparsity assumption for domain-specific topics $t \in S$: distributions ϕ_{wt}, θ_{td} contain many zero probabilities.

Maximize the sum of KL-divergences $\mathrm{KL}(\beta \| \phi_t)$ and $\mathrm{KL}(\alpha \| \theta_d)$:

$$R(\Phi, \Theta) = -\beta_0 \sum_{t \in S} \sum_{w \in W} \beta_w \ln \phi_{wt} - \alpha_0 \sum_{d \in D} \sum_{t \in S} \alpha_t \ln \theta_{td} \to \max .$$

The regularized M-step gives "anti-LDA", for all $t \in S$:

$$\phi_{wt} \propto \big(n_{wt} - \beta_0 \beta_w\big)_+, \qquad \theta_{td} \propto \big(n_{td} - \alpha_0 \alpha_t\big)_+ .$$

Varadarajan J., Emonet R., Odobez J.-M. A sparsity constraint for topic models — application to temporal activity mining // NIPS-2010 Workshop on Practical Applications of Sparse Modeling: Open Issues and New Directions.

Regularization for topics decorrelation

The dissimilarity assumption:
domain-specific topics $t \in S$ must be as distant as possible.

Maximize covariances between column vectors ϕ_t:

$$R(\Phi) = -\frac{\tau}{2} \sum_{t,s \in S} \sum_{w \in W} \phi_{wt} \phi_{ws} \to \max .$$

The regularized M-step makes columns of Φ more distant:

$$\phi_{wt} \propto \Bigl(n_{wt} - \tau \phi_{wt} \sum_{s \in S \backslash t} \phi_{ws} \Bigr)_{+}.$$

Tan Y., Ou Z. Topic-weak-correlated latent Dirichlet allocation // 7th Int'l Symp. Chinese Spoken Language Processing (ISCSLP), 2010. — Pp. 224–228.

Regularization for topic selection

Assumption: infrequent topics are not well-interpretable.

Maximize KL-divergence $\mathrm{KL}\big(\frac{1}{|T|} \,\|\, p(t)\big)$ to make distribution over topics $p(t) = \sum_d p(d)\theta_{td}$ sparse:

$$R(\Theta) = -\tau \sum_{t \in S} \ln \sum_{d \in D} p(d)\theta_{td} \to \max.$$

The regularized M-step formula results in Θ rows sparsing:

$$\theta_{td} \propto \left(n_{td} - \tau \frac{n_d}{n_t} \theta_{td}\right)_+.$$

Effect: if n_t is small then in the t-th row may turn into zeros.

Vorontsov K. V., Potapenko A. A., Plavin A. V. Additive regularization of topic models for topic selection and sparse factorization // SLDS 2015, Royal Holloway, University of London, UK. pp. 193–202.

Combining topic models by adding their regularizers

Maximum log-likelihood with additive combination of regularizers:

$$\sum_{d,w} n_{dw} \ln \sum_t \phi_{wt}\theta_{td} + \sum_{i=1}^{n} \tau_i R_i(\Phi, \Theta) \to \max_{\Phi,\Theta},$$

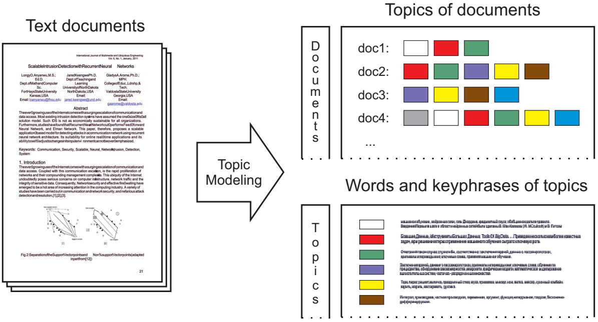
where τ_i are regularization coefficients.

EM-algorithm is a simple iteration method for the system

$$\begin{array}{rl} \text{E-step:} & \left\{ \begin{array}{l} p_{tdw} = \operatorname*{norm}_{t\in T}\big(\phi_{wt}\theta_{td}\big) \\ \phi_{wt} = \operatorname*{norm}_{w\in W}\Big(\sum_{d\in D} n_{dw}p_{tdw} + \phi_{wt}\sum_{i=1}^{n}\tau_i\frac{\partial R_i}{\partial \phi_{wt}}\Big) \\ \theta_{td} = \operatorname*{norm}_{t\in T}\Big(\sum_{w\in d} n_{dw}p_{tdw} + \theta_{td}\sum_{i=1}^{n}\tau_i\frac{\partial R_i}{\partial \theta_{td}}\Big) \end{array} \right. \\ \text{M-step:} & \end{array}$$

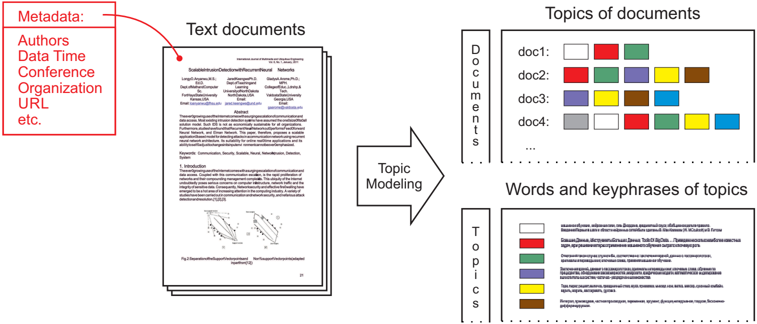
Multimodal Probabilistic Topic Modeling

Given a text document collection *Probabilistic Topic Model* finds:
$p(t|d)$ — topic distribution for each document d,
$p(w|t)$ — term distribution for each topic t.



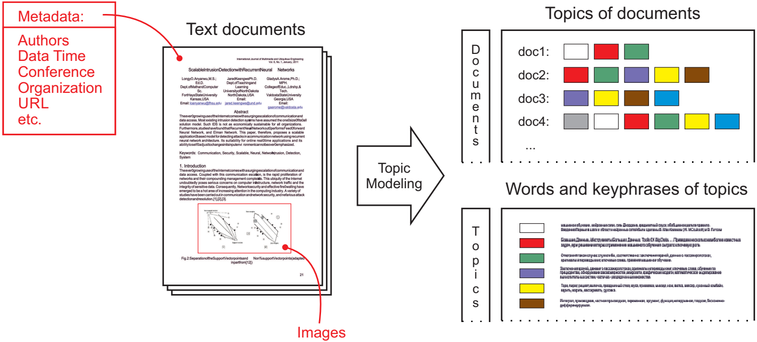
Multimodal Probabilistic Topic Modeling

Multimodal Topic Model finds topical distribution for terms $p(w|t)$, authors $p(a|t)$, time $p(y|t)$,



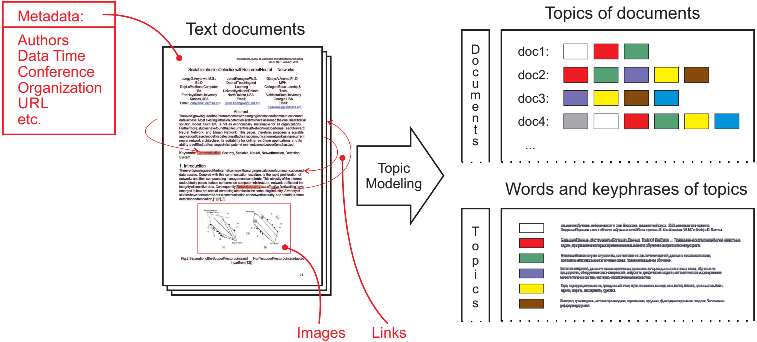
Multimodal Probabilistic Topic Modeling

Multimodal Topic Model finds topical distribution for terms $p(w|t)$, authors $p(a|t)$, time $p(y|t)$, objects on images $p(o|t)$,



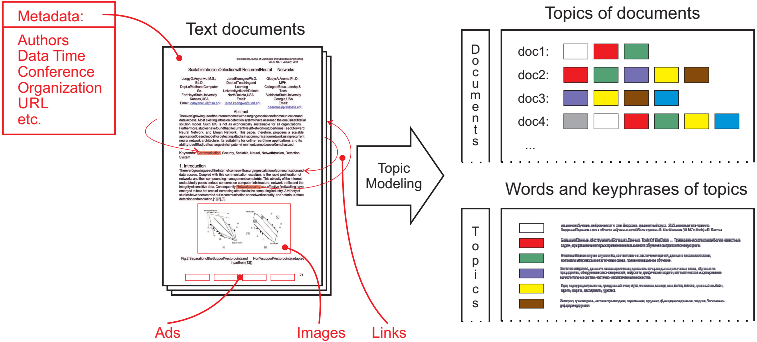
Multimodal Probabilistic Topic Modeling

Multimodal Topic Model finds topical distribution for terms $p(w|t)$, authors $p(a|t)$, time $p(y|t)$, objects on images $p(o|t)$, linked documents $p(d'|t)$,



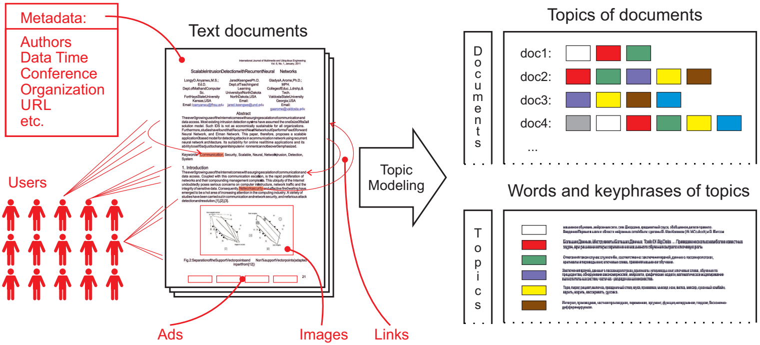
Multimodal Probabilistic Topic Modeling

Multimodal Topic Model finds topical distribution for terms $p(w|t)$, authors $p(a|t)$, time $p(y|t)$, objects on images $p(o|t)$, linked documents $p(d'|t)$, advertising banners $p(b|t)$,



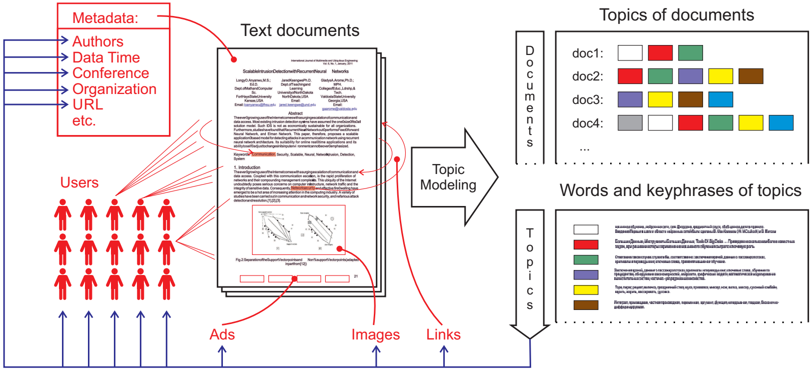
Multimodal Probabilistic Topic Modeling

Multimodal Topic Model finds topical distribution for terms $p(w|t)$, authors $p(a|t)$, time $p(y|t)$, objects on images $p(o|t)$, linked documents $p(d'|t)$, advertising banners $p(b|t)$, users $p(u|t)$,



Multimodal Probabilistic Topic Modeling

Multimodal Topic Model finds topical distribution for terms $p(w|t)$, authors $p(a|t)$, time $p(y|t)$, objects on images $p(o|t)$, linked documents $p(d'|t)$, advertising banners $p(b|t)$, users $p(u|t)$, and binds all these modalities into a single topic model.



Multimodal extension of ARTM [Vorontsov, 2015]

W^m is a vocabulary of tokens of m-th modality, $m \in M$
$W = W^1 \sqcup \dots \sqcup W^M$ is a joint vocabulary of all modalities

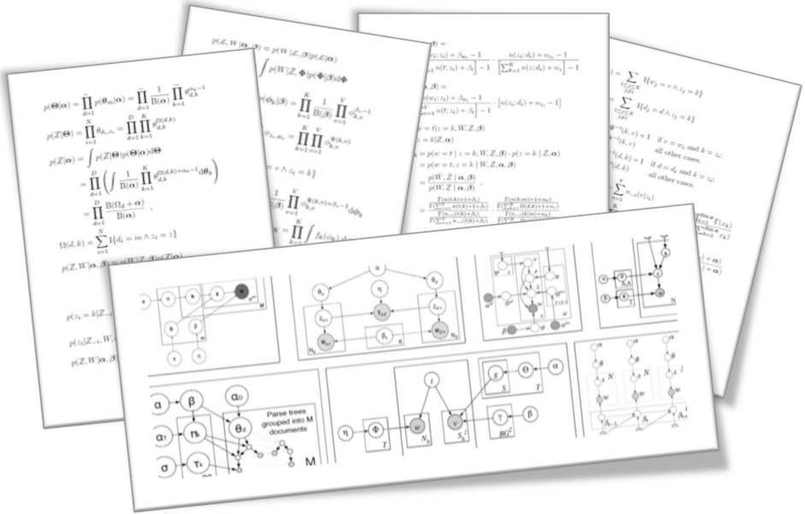
Maximum multimodal log-likelihood with regularization:

$$\sum_{m \in M} \lambda_m \sum_{d \in D} \sum_{w \in W^m} n_{dw} \ln \sum_t \phi_{wt} \theta_{td} + R(\Phi, \Theta) \to \max_{\Phi, \Theta}$$

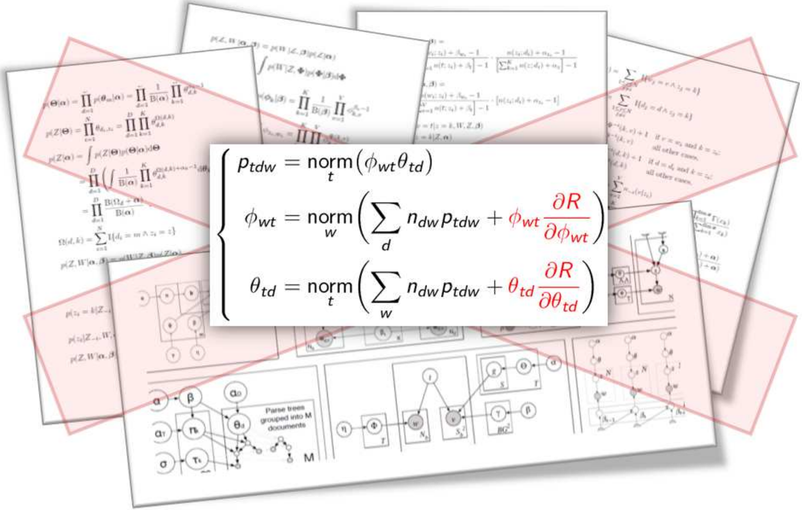
EM-algorithm is a simple iteration method for the system

$$\begin{array}{ll} \text{E-step:} & \left\{ \begin{array}{l} p_{tdw} = \operatorname*{norm}_{t \in T} \big(\phi_{wt} \theta_{td} \big) \\ \phi_{wt} = \operatorname*{norm}_{w \in W^m} \Big(\sum_{d \in D} \lambda_{m(w)} n_{dw} p_{tdw} + \phi_{wt} \frac{\partial R}{\partial \phi_{wt}} \Big) \\ \theta_{td} = \operatorname*{norm}_{t \in T} \Big(\sum_{w \in d} \lambda_{m(w)} n_{dw} p_{tdw} + \theta_{td} \frac{\partial R}{\partial \theta_{td}} \Big) \end{array} \right. \\ \text{M-step:} & \end{array}$$

Bayesian learning is a too complicated theory for PTM



ARTM provides easier understanding and combining of PTMs



$$\begin{cases} p_{tdw} = \underset{t}{\operatorname{norm}}\big(\phi_{wt}\theta_{td}\big) \\ \phi_{wt} = \underset{w}{\operatorname{norm}}\Big(\sum_{d} n_{dw}p_{tdw} + \phi_{wt}\frac{\partial R}{\partial \phi_{wt}}\Big) \\ \theta_{td} = \underset{t}{\operatorname{norm}}\Big(\sum_{w} n_{dw}p_{tdw} + \theta_{td}\frac{\partial R}{\partial \theta_{td}}\Big) \end{cases}$$

BigARTM project

BigARTM features:

- Parallel + Online + Multimodal + Regularized Topic Modeling
- Out-of-core one-pass processing of Big Data
- Built-in library of regularizers and quality measures

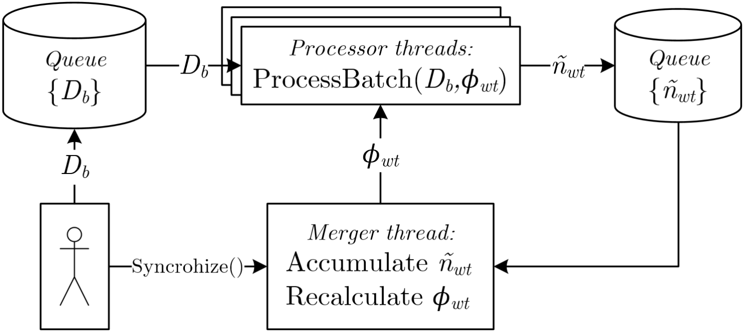
BigARTM community:

- 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

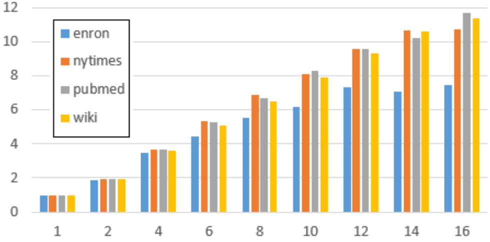


- Concurrent processing of batches $D = D_1 \sqcup \dots \sqcup D_B$
- Simple single-threaded code for *ProcessBatch*
- User controls when to update the model in online algorithm
- Deterministic (reproducible) results from run to run

Experiment 1: 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 2: BigARTM vs Gensim vs Vowpal Wabbit

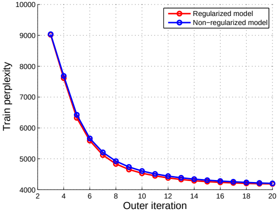
- 3.7M articles from Wikipedia, 100K unique words

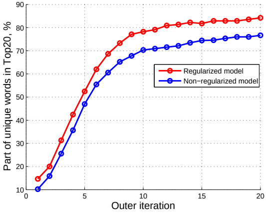
	procs	train	inference	perplexity
BigARTM	1	35 min	72 sec	4000
Gensim.LdaModel	1	369 min	395 sec	4161
VowpalWabbit.LDA	1	73 min	120 sec	4108
BigARTM	4	9 min	20 sec	4061
Gensim.LdaMulticore	4	60 min	222 sec	4111
BigARTM	8	4.5 min	14 sec	4304
Gensim.LdaMulticore	8	57 min	224 sec	4455

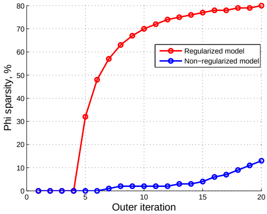
- *procs* = number of parallel threads
- *inference* = time to infer θ_d for 100K held-out documents
- *perplexity* is calculated on held-out documents.

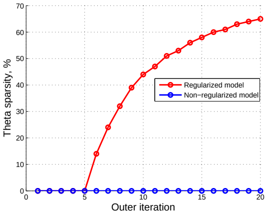
Experiment 3: Running BigARTM with multiple regularizers

ARTM combines regularizers to improve sparsity and the number of topical words without a loss of the perplexity.



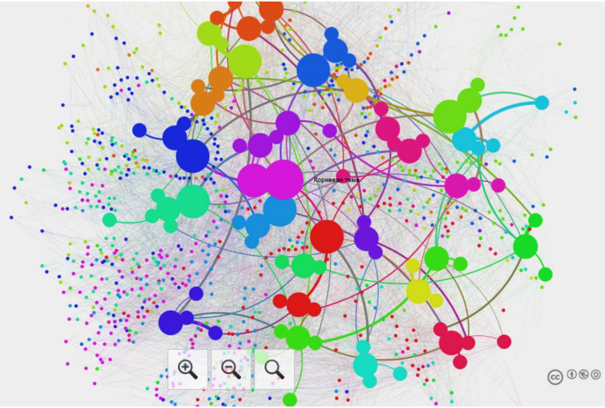






Experiment 4: Hierarchical topic model for MMPR-IIP conferences

$|D| = 865$, $|W| = 42\,000$ n-grams, in Russian
BigARTM is used with 7 regularizers to build 3-level hierarchy.



http://explore-mmro.ru

Experiment 5: The interpretability of n-gram models

Two modalities — unigrams & bigrams

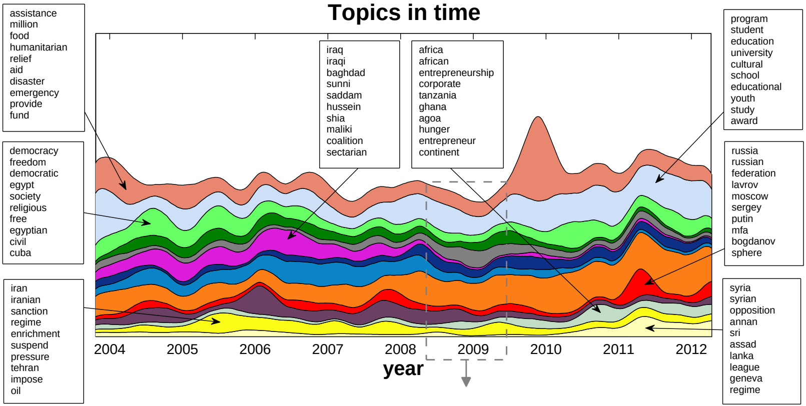
MMPR-IIP conferences collection, $|D| = 865$, in Russian

pattern recognition in bioinformatics		optimization and computational complexity	
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

Experiment 6. Temporal topic model of political press-releases

20 000 press-releases from 2003 to 2013, 180Mb.

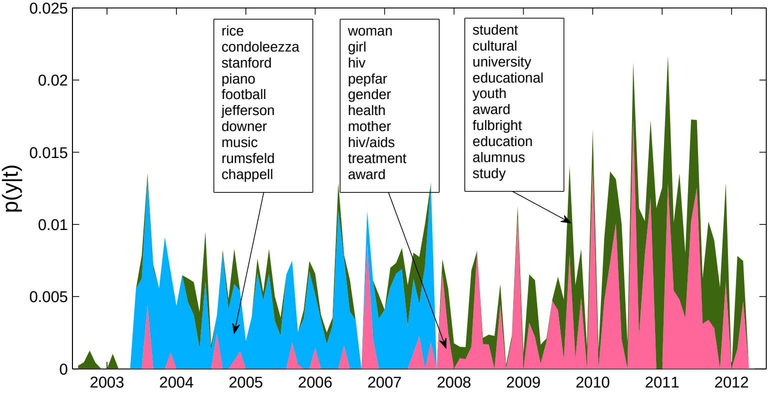
Examples of most valuable topics



Experiment 6. Temporal topic model of political press-releases

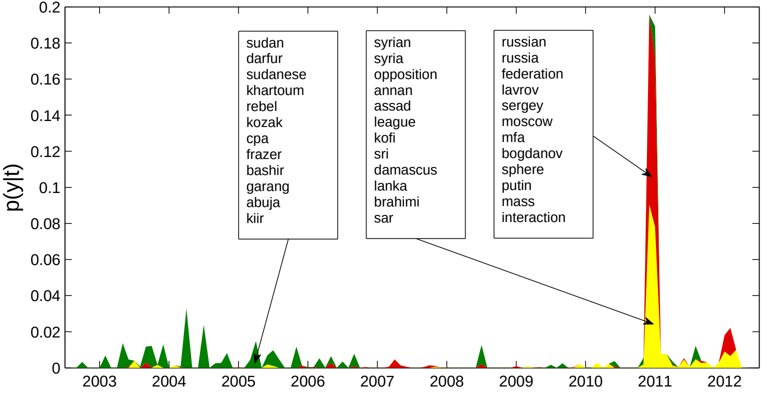
20 000 press-releases from 2003 to 2013, 180Mb.

Examples of permanent topics



Experiment 6. Temporal topic model of political press-releases

20 000 press-releases from 2003 to 2013, 180Mb.
Examples of event topics



Brief summary

- **Exploratory Search:** a paradigm of Information Retrieval for professionals, researchers, students, and inquisitive persons
- **Multi-criteria Topic Modeling:** a way to meet multiple requirements coming from Exploratory Search
- **ARTM:** a novel non-Bayesian approach for multi-criteria optimization and combining Topic Models
- **BigARTM:** open source project for parallel online multimodal **A**dditively **R**egularized **T**opic **M**odeling of large collections

http://bigartm.org • Join BigARTM community!

References

- *Hofmann T.* Probabilistic Latent Semantic Indexing. ACM SIGIR, 1999.
- *Blei D., Ng A., Jordan M.* Latent Dirichlet Allocation. Journal of Machine Learning Research, 2003. No. 3, pp. 993–1022.
- *Asuncion A., Welling M., Smyth P., Teh Y. W.* On smoothing and inference for topic models. Int'l Conf. on Uncertainty in Artificial Intelligence, 2009.
- *Vorontsov K. V., Potapenko A. A.* Tutorial on Probabilistic Topic Modeling: Additive Regularization for Stochastic Matrix Factorization. AIST'2014, Analysis of Images, Social networks and Texts. Springer, 2014. CCIS, Vol. 436. pp. 29–46.
- *Vorontsov K. V., Frei O. I., Apishev M. A., Romov P. A., Suvorova M. A., Yanina A. O.* Non-Bayesian Additive Regularization for Multimodal Topic Modeling of Large Collections. Topic Models: Post-Processing and Applications, CIKM 2015 Workshop, October 19, 2015, Melbourne, Australia.