

# Аддитивная регуляризация тематических моделей связного текста

Воронцов Константин Вячеславович

(МФТИ • ФИЦ ИУ РАН • МГУ • Яндекс • FORECSYS • Aithea)



Таганрог • 9–13 октября 2017

- 1 Вероятностное тематическое моделирование**
  - Задача тематического моделирования
  - Аддитивная регуляризация
  - Модели связного текста
- 2 Тематические модели связного текста**
  - Мультиграммные модели
  - Модели совстречаемости слов
  - Модели тематической сегментации
- 3 Разведочный информационный поиск**
  - Задача разведочного информационного поиска
  - Выбор траектории регуляризации
  - Эксперименты

## Приложения тематического моделирования

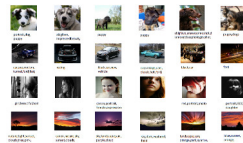
разведочный поиск в  
электронных библиотеках



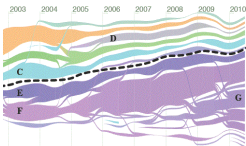
персонализированный  
поиск в соцсетях



мультимодальный поиск  
текстов и изображений



детектирование и трекинг  
новостных сюжетов



навигация по большим  
текстовым коллекциям



управление диалогом в  
разговорном интеллекте



## Пусть

- $W$  — конечное множество слов (терминов, токенов)
- $D$  — конечное множество текстовых документов
- $T$  — конечное множество тем
- каждое слово  $w$  в документе  $d$  связано с некоторой темой  $t$
- $D \times W \times T$  — дискретное вероятностное пространство
- **порядок слов в документе не важен (bag of words)**
- порядок документов в коллекции не важен
- коллекция — это i.i.d. выборка  $(d_i, w_i, t_i)_{i=1}^n \sim p(d, w, t)$
- $d_i, w_i$  — наблюдаемые, темы  $t_i$  — скрытые
- гипотеза условной независимости:  $p(w|d, t) = p(w|t)$

Тематическая модель, по формуле полной вероятности:

$$p(w|d) = \sum_{t \in T} p(w|t) p(t|d)$$

## Постановка задачи тематического моделирования

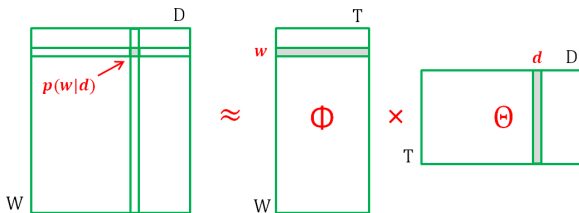
**Дано:** коллекция текстовых документов

- $n_{dw}$  — частоты терминов в документах,  $\hat{p}(w|d) = \frac{n_{dw}}{n_d}$

**Найти:** параметры тематической модели  $p(w|d) = \sum_{t \in T} \phi_{wt} \theta_{td}$

- $\phi_{wt} = p(w|t)$  — вероятности терминов  $w$  в каждой теме  $t$
- $\theta_{td} = p(t|d)$  — вероятности тем  $t$  в каждом документе  $d$

Это задача стохастического матричного разложения:



## Задачи, некорректно поставленные по Адамару

Задача *корректно поставлена*,  
если её решение

- существует,
- единственно,
- устойчиво.



Жак Саломон Адамар  
(1865–1963)

Наша задача матричного разложения *некорректно поставлена*:  
если  $\Phi, \Theta$  — решение, то стохастические  $\Phi', \Theta'$  — тоже решения

- $\Phi'\Theta' = (\Phi S)(S^{-1}\Theta)$ ,  $\text{rank } S = |T|$
- $\mathcal{L}(\Phi', \Theta') = \mathcal{L}(\Phi, \Theta)$
- $\mathcal{L}(\Phi', \Theta') \leq \mathcal{L}(\Phi, \Theta) + \varepsilon$  — приближённые решения

**Регуляризация** — стандартный приём доопределения решения  
с помощью дополнительных критериев.

## ARTM — Аддитивная Регуляризация Тематических Моделей

Максимизация  $\log$  правдоподобия с регуляризатором  $R$ :

$$\sum_{d,w} n_{dw} \ln \sum_{t \in T} \phi_{wt} \theta_{td} + R(\Phi, \Theta) \rightarrow \max_{\Phi, \Theta}; \quad R(\Phi, \Theta) = \sum_i \tau_i R_i(\Phi, \Theta)$$

EM-алгоритм: метод простой итерации для системы уравнений

$$\begin{array}{l} \text{E-шаг:} \\ \text{M-шаг:} \end{array} \left\{ \begin{array}{l} p(t|d, w) = \mathop{\text{norm}}_{t \in T}(\phi_{wt} \theta_{td}) \\ \phi_{wt} = \mathop{\text{norm}}_{w \in W} \left( \sum_{d \in D} n_{dw} p(t|d, w) + \phi_{wt} \frac{\partial R}{\partial \phi_{wt}} \right) \\ \theta_{td} = \mathop{\text{norm}}_{t \in T} \left( \sum_{w \in W} n_{dw} p(t|d, w) + \theta_{td} \frac{\partial R}{\partial \theta_{td}} \right) \end{array} \right.$$

где  $\mathop{\text{norm}}_{t \in T}(x_t) = \frac{\max\{x_t, 0\}}{\sum_{s \in T} \max\{x_s, 0\}}$  — операция нормировки вектора.

Воронцов К. В. Аддитивная регуляризация тематических моделей коллекций текстовых документов. Доклады РАН. 2014.

## Мультимодальная ARTM

$W^m$  — словарь токенов  $m$ -й модальности,  $m \in M$

Максимизация суммы  $\log$  правдоподобий с регуляризацией:

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

EM-алгоритм: метод простой итерации для системы уравнений

$$\begin{array}{l} \text{E-шаг:} \\ \text{M-шаг:} \end{array} \left\{ \begin{array}{l} p(t|d, w) = \operatorname{norm}_{t \in T}(\phi_{wt} \theta_{td}) \\ \phi_{wt} = \operatorname{norm}_{w \in W^m} \left( \sum_{d \in D} \tau_{m(w)} n_{dw} p(t|d, w) + \phi_{wt} \frac{\partial R}{\partial \phi_{wt}} \right) \\ \theta_{td} = \operatorname{norm}_{t \in T} \left( \sum_{w \in W^d} \tau_{m(w)} n_{dw} p(t|d, w) + \theta_{td} \frac{\partial R}{\partial \theta_{td}} \right) \end{array} \right.$$

*K.Vorontsov, O.Frei, M.Apishev et al.* Non-bayesian additive regularization for multimodal topic modeling of large collections. CIKM TM workshop, 2015.



## BigARTM: библиотека тематического моделирования

### Ключевые возможности:

- Онлайн-овый параллельный мультимодальный ARTM
- Большие данные: коллекция не хранится в памяти
- Встроенная библиотека регуляризаторов и мер качества

### Сообщество:

- Открытый код <https://github.com/bigartm>  
(discussion group, issue tracker, pull requests)
- Документация <http://bigartm.org>



### Лицензия и среда разработки:

- 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

## BigARTM упрощает разработку тематических моделей


Для построения сложных моделей в BigARTM не нужны ни математические выкладки, ни программирование «с нуля».


### Этапы моделирования

### Bayesian TM

### ARTM

	Bayesian TM	ARTM	
	Анализ требований	Анализ требований	
Формализация:	Вероятностная порождающая модель данных	Стандартные критерии	Свои критерии
Алгоритмизация:	Байесовский вывод для данной порождающей модели (VI, GS, EP)	Общий регуляризованный EM-алгоритм для любых моделей	
Реализация:	Исследовательский код (Matlab, Python, R)	Промышленный код BigARTM (C++, Python API)	
Оценивание:	Исследовательские метрики, исследовательский код	Стандартные метрики	Свои метрики
	Внедрение	Внедрение	

 -- нестандартизируемые этапы, уникальная разработка для каждой задачи

 -- стандартизуемые этапы

## Тематические модели связного текста (beyond bag-of-words)

n-gram



Модели с модальностями  $n$ -грамм, коллокаций, именованных сущностей

syntax



Модели, учитывающие результаты автоматического синтаксического разбора (SyntaxNet)

coherence



Модели дистрибутивной семантики на основе совстречаемости слов (битермы, когерентность)

sentence



Тематические модели, учитывающие границы предложений, абзацев и секций документов

segmentation



Тематические модели сегментации с автоматическим определением границ сегментов

## Биграммы радикально улучшают интерпретируемость тем

Коллекция 1000 статей конференций ММРО, ИОИ на русском

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

*Сергей Стенин. Мультиграммные аддитивно регуляризованные тематические модели. Магистерская диссертация, МФТИ, 2015.*

## Задача автоматического выделения терминов

*Термин* — фраза ( $n$ -грамма) со следующим набором свойств:

- 1 *высокая частотность* (frequency):  
много раз встречается в коллекции;
- 2 *совстречаемость слов* (collocation):  
состоит из слов, неслучайно часто встречающихся вместе;
- 3 *полнота* (completeness):  
является максимальной по включению цепочкой слов;
- 4 *синтаксическая связность* (syntactic connectedness):  
является грамматически корректным словосочетанием;
- 5 *тематичность* (topicality):  
часто встречается в небольшом числе тем.

Сумма технологий для АТЕ (Authomatic Term Extraction):

TopMine ① ② ③ + SyntaxNet ④ + BigARTM ⑤

## Алгоритм TopMine: определения и основные идеи

- Хэш-таблица  $C(a_1, \dots, a_k)$  счётчиков частых  $k$ -грамм, инициализируется для всех униграмм  $a$  с частотой  $n_a \geq \varepsilon_1$
- Свойство антимонотонности:

$$C(a_1, \dots, a_k) \geq C(a_1, \dots, a_k, a_{k+1}).$$

- $A_{d,k}$  — множество позиций  $i$  в документе  $d$  таких, что

$$C(w_{d,i}, \dots, w_{d,i+k-1}) \geq \varepsilon_k,$$

инициализируется для всех частых униграмм.

- Основной шаг алгоритма: для всех  $i = 1, \dots, n_d$   
**если**  $(i \in A_{d,k})$  **и**  $(i + 1 \in A_{d,k})$  **то**  $++C(w_{d,i}, \dots, w_{d,i+k})$ .

---

Ahmed El-Kishky, Yanglei Song, Chi Wang, Clare R. Voss, Jiawei Han.  
Scalable Topical Phrase Mining from Text Corpora // VLDB, 2015.

## Алгоритм TopMine: быстрый поиск высокочастотных $k$ -грамм

**Вход:** коллекция  $D$ , пороги  $\varepsilon_k$ ;

**Выход:** хэш-таблица частот  $C(a_1, \dots, a_k)$ ,  $k = 1, \dots, k_{\max}$ ;

$A_{d,1} := \{1, \dots, n_d\}$ ;

$C(w) := n_w$  для всех  $w \in W$  таких, что  $n_w \geq \varepsilon_1$ ;

**для**  $k := 2, \dots, k_{\max}$  **пока**  $D \neq \emptyset$

**для всех**  $d \in D$

$A_{d,k} := \{i \in A_{d,k-1} \mid C(w_{d,i}, \dots, w_{d,i+k-2}) \geq \varepsilon_k\}$ ;

**если**  $A_{d,k} = \emptyset$  **то**  $D := D \setminus \{d\}$ ;

**для всех**  $i \in A_{d,k}$

**если**  $i+1 \in A_{d,k}$  **то**  $++C(w_{d,i}, \dots, w_{d,i+k-1})$ ;

    оставить только частые  $k$ -граммы:  $C(a_1, \dots, a_k) \geq \varepsilon_k$ ;

Преимущество алгоритма: линейная память и скорость.

Ahmed El-Kishky, Yanglei Song, Chi Wang, Clare R. Voss, Jiawei Han.  
Scalable Topical Phrase Mining from Text Corpora // VLDB, 2015.

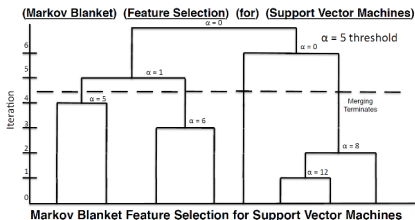
## Алгоритм TopMine: отбор фраз по встречаемости и полноте

Итеративное слияние фраз с понижением значимости  $\alpha$ .

$p_u$  — оценка вероятности встретить фразу  $u$

$p_{uv}$  — оценка вероятности встретить фразу  $uv$

Критерии:  $\text{SignificanceScore} = \frac{p_{uv} - p_u p_v}{\sqrt{p_{uv}}}$  или  $\text{PMI} = \log \frac{p_{uv}}{p_u p_v}$



Ahmed El-Kishky, Yanglei Song, Chi Wang, Clare R. Voss, Jiawei Han.  
 Scalable Topical Phrase Mining from Text Corpora // VLDB, 2015.



## Синтаксический анализатор Google SyntaxNet

SyntaxNet — предобученная нейросеть поверх TensorFlow, поддерживает 40 языков, включая русский.

### Вход:

- список предложений

Выход, для каждого слова в каждом предложении:

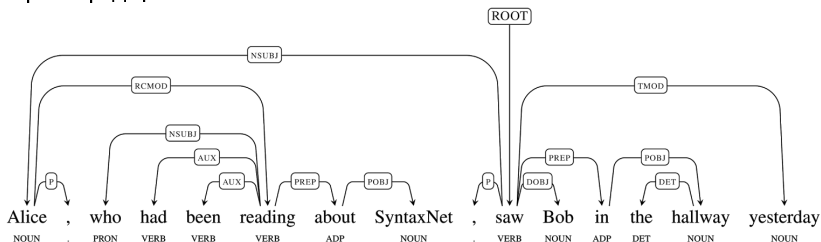
- id (порядковый номер слова в предложении)
- id родительского слова (0 для корня)
- исходное слово
- нормальная форма
- часть речи: NOUN, VERB, ADJ, ADV, ...
- член предложения: nsubj, dobj, conj, cc, nmod, ...

---

*D.Andor, C.Alberti, D.dWeiss, A.Severyn, A.Presta, K.Ganchev, S.Petrov, M.Collins.* Globally Normalized Transition-Based Neural Networks. 2016.

## Синтаксический анализатор Google SyntaxNet

Пример дерева зависимостей:



Варианты стратегий отбора терминов-кандидатов:

- брать все поддеревья
- брать все именные группы (корень — NOUN)
- не брать CONJ, SCONJ, DET, AUX, INTJ, PART, PUNCT, SYM

Announcing SyntaxNet: the world's most accurate parser goes open source.  
<https://research.googleblog.com/2016/05/announcing-syntaxnet-worlds-most.html>.

## Постановка эксперимента

- Коллекция  $|D| = 3200$  аннотаций статей NIPS (Neural Information Processing Systems),  $n = 500\,000$  слов
- Ручная разметка небольшого *случайного* подмножества (2000  $n$ -грамм) на термины / не-термины
- Train : Test = 1000 : 1000
- 7 статистических признаков из TopMine
- 2 синтаксических признака из SyntaxNet
- 3 тематических признака из BigARTM, 30 тем
- две модели классификации:  
логистическая регрессия, градиентный бустинг

---

*Владимир Полушин.* Тематические модели для ранжирования рекомендаций текстового контента. Бакалаврская диссертация, ВМК МГУ, 2017.

## Сравнение методов автоматического отбора терминов

Найти *как можно больше терминов* — полнота важнее точности

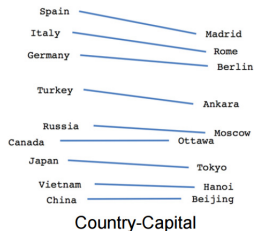
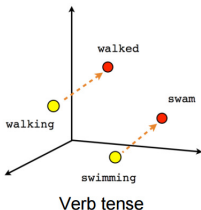
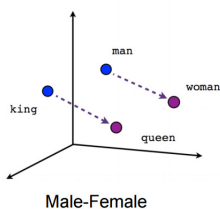
Группа признаков			Линейная модель			Градиентный бустинг		
Синт	Стат	Тем	AUC	Точность	Полнота	AUC	Точность	Полнота
+			0.83	0.20	0.91	0.83	0.20	0.91
	+		0.71	0.09	0.94	0.73	0.11	0.90
		+	0.92	0.32	1.00	0.95	0.32	1.00
+	+		0.88	0.22	0.91	0.88	0.24	0.91
+		+	0.91	0.36	0.91	0.95	0.34	0.99
	+	+	0.93	0.29	0.94	0.98	0.34	1.00
+	+	+	<b>0.95</b>	<b>0.38</b>	<b>0.91</b>	<b>0.97</b>	<b>0.41</b>	<b>0.99</b>

$$\boxed{\text{Стат}} < \boxed{\text{Син}} < \boxed{\text{Син}+\text{Стат}} < \boxed{\text{Тем}} < \boxed{\begin{matrix} \text{Стат}+\text{Тем} \\ \text{Син}+\text{Тем} \end{matrix}} < \boxed{\text{Стат}+\text{Син}+\text{Тем}}$$

- Тематические признаки существенно повышают качество
- Синтаксические признаки можно не использовать

## Задача семантического векторного представления слов

Найти для каждого слова  $w$  вектор  $v_w \in \mathbb{R}^T$ , чтобы близкие по смыслу слова имели близкие векторы.



### Дистрибутивная гипотеза

- Words that occur in the same contexts tend to have similar meanings [Harris, 1954].
- You shall know a word by the company it keeps [Firth, 1957].

## Формализация дистрибутивной гипотезы в программе word2vec

**Дано:**  $n_{uw}$  — совстречаемость слов  $u, w$  в окне  $\pm h$  слов

**Найти:** семантические векторные представления слов  $v_w$

**Модель:** вероятность слова  $w$  в контексте слова  $u$ , то есть при условии, что слово  $u$  находится рядом:

$$p(w|u) = \text{SoftMax}_{w \in W} \langle v_w, v_u \rangle = \frac{\exp \langle v_w, v_u \rangle}{\sum_v \exp \langle v_v, v_u \rangle}$$

**Критерий:** максимум log-правдоподобия:

$$\sum_{w, u \in W} n_{wu} \ln p(w|u) \rightarrow \max_{\{v_w\}}$$

**Проблема:** координаты векторов не интерпретируемы

---

*T. Mikolov, K. Chen, G. Corrado, J. Dean.* Efficient estimation of word representations in vector space. CoRR, 2013.

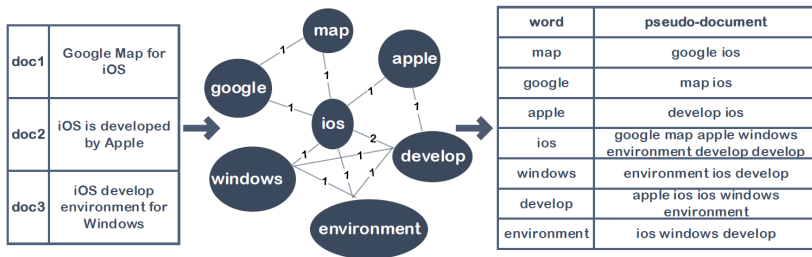
## Модель сети слов WNTM для коротких текстов

**Идея:** моделировать не документы, а связи между словами.

$d_w$  — псевдо-документ, объединение всех контекстов слова  $w$ .

$n_{wu}$  — число вхождений слова  $u$  в псевдо-документ  $d_w$ .

**Контекст** — короткое сообщение / предложение / окно  $\pm h$  слов.



Yuan Zuo, Jichang Zhao, Ke Xu. **Word Network Topic Model**: a simple but general solution for short and imbalanced texts. 2014.

## Модель битермов $\text{BitermTM}$ и модель сети слов $\text{WNTM}$

*Битерм* — пара слов, встречающихся рядом:  
в одном коротком сообщении / предложении / окне  $\pm h$  слов.  
Регуляризатор, эквивалентный модели битермов  $\text{BitermTM}$ :

$$R(\Phi) = \tau \sum_{u,v \in W} n_{uv} \ln \sum_{t \in T} n_t \phi_{ut} \phi_{vt} \rightarrow \max$$

Регуляризатор, эквивалентный модели сети слов  $\text{WNTM}$ :

$$R(\Phi, \Theta') = \tau \sum_{u,v \in W} n_{uv} \ln \sum_{t \in T} \phi_{ut} \theta'_{tv} \rightarrow \max_{\Phi, \Theta'}$$

---

*Xiaohui Yan, Jiafeng Guo, Yanyan Lan, Xueqi Cheng.* A **Biterm Topic Model** for short texts. WWW, 2013.

*Yuan Zuo, Jichang Zhao, Ke Xu.* **Word Network Topic Model**: a simple but general solution for short and imbalanced texts. 2014.

*Berlin Chen.* **Word Topic Models** for spoken document retrieval and transcription. ACM Trans., 2009.



## Примеры векторных операций в задаче аналогии слов

Два подхода к синтезу векторных представлений слов:

- **ARTM**: интерпретируемые разреженные компоненты
- **word2vec**: интерпретируемые векторные операции

Операция	Результат ARTM	Результат word2vec
king – boy + girl	<i>queen, princess, lord, prince</i>	<i>queen, princess, regnant, kings</i>
moscow – russia + spain	<i>madrid, barcelona, aires, buenos</i>	<i>madrid, barcelona, valladolid, malaga</i>
india – russia + ruble	<i>rupee, birbhum, pradesh, madhaya</i>	<i>rupee, rupiah, devalued, debased</i>
cars – car + computer	<i>computers, software, servers, implementations</i>	<i>computers, software, hardware, microcomputers</i>

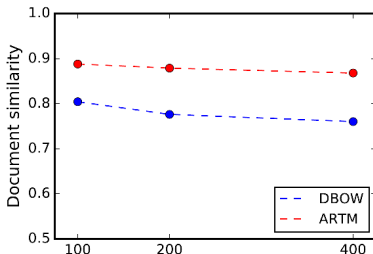
---

*A.Potapenko, A.Popov, K.Vorontsov. Interpretable probabilistic embeddings: bridging the gap between topic models and neural networks. AINL-6, 2017.*

## Задача семантической близости документов

Коллекция 20 000 триплетов статей arXiv [Dai, 2015]:  
(статья А, похожая статья В, не похожая статья С)

- Обучение моделей по выборке 1М статей arXiv
- Эталон для сравнения: DBOW paragraph2vec [Dai, 2015]



Andrew Dai, Cristopher Olah, Quoc Le. Document Embedding with Paragraph Vectors. CoRR, 2015.

A.Potapenko, A.Popov, K.Vorontsov. Interpretable probabilistic embeddings: bridging the gap between topic models and neural networks. AINL-6, 2017.

## Задача тематической сегментации документов

Документ  $d = \{w_1, \dots, w_{n_d}\}$ ,  $n_d$  — длина документа  $d$

Матрица тематики слов в документах  $p(t|d, w_i)$  размера  $T \times n_d$ :



## Регуляризация (пост-обработка) E-шага

Трёхмерная матрица  $\Pi = (p_{tdw} = p(t|d, w))_{T \times D \times W}$

Максимизация  $\log$  правдоподобия с регуляризаторами  $R$  и  $\tilde{R}$ :

$$\sum_{d \in D} \sum_{w \in d} n_{dw} \ln \sum_{t \in T} \phi_{wt} \theta_{td} + R(\Pi(\Phi, \Theta)) + \tilde{R}(\Phi, \Theta) \rightarrow \max_{\Phi, \Theta}$$

EM-алгоритм: метод простой итерации для системы уравнений

$$\begin{cases} \text{E-шаг:} & \left\{ \begin{array}{l} p_{tdw} = \operatorname{norm}_{t \in T}(\phi_{wt} \theta_{td}) \\ \tilde{p}_{tdw} = p_{tdw} \left( 1 + \frac{1}{n_{dw}} \left( \frac{\partial R(\Pi)}{\partial p_{tdw}} - \sum_{z \in T} p_{zdw} \frac{\partial R(\Pi)}{\partial p_{zdw}} \right) \right) \end{array} \right. \\ \text{M-шаг:} & \left\{ \begin{array}{l} \phi_{wt} = \operatorname{norm}_{w \in W} \left( \sum_{d \in D} n_{dw} \tilde{p}_{tdw} + \phi_{wt} \frac{\partial \tilde{R}}{\partial \phi_{wt}} \right) \\ \theta_{td} = \operatorname{norm}_{t \in T} \left( \sum_{w \in d} n_{dw} \tilde{p}_{tdw} + \theta_{td} \frac{\partial \tilde{R}}{\partial \theta_{td}} \right) \end{array} \right. \end{cases}$$

## Доказательство

**Лемма 1.** Для функции  $p_{tdw}(\Phi, \Theta) = \frac{\phi_{wt}\theta_{td}}{\sum_z \phi_{wz}\theta_{zd}}$  и любого  $z \in T$

$$\phi_{wt} \frac{\partial p_{zdw}}{\partial \phi_{wt}} = \theta_{td} \frac{\partial p_{zdw}}{\partial \theta_{td}} = p_{tdw} ([z=t] - p_{zdw}).$$

Введём функцию от вспомогательных переменных  $\Pi$ :

$$Q_{tdw}(\Pi) = \frac{\partial R(\Pi)}{\partial p_{tdw}} - \sum_{z \in T} p_{zdw} \frac{\partial R(\Pi)}{\partial p_{zdw}}.$$

**Лемма 2.** Если  $R(\Pi)$  не зависит от  $p_{tdw}$  при  $w \notin d$ , то

$$\phi_{wt} \frac{\partial R(\Pi)}{\partial \phi_{wt}} = \sum_{d \in D} p_{tdw} Q_{tdw}(\Pi); \quad \theta_{td} \frac{\partial R(\Pi)}{\partial \theta_{td}} = \sum_{w \in D} p_{tdw} Q_{tdw}(\Pi).$$

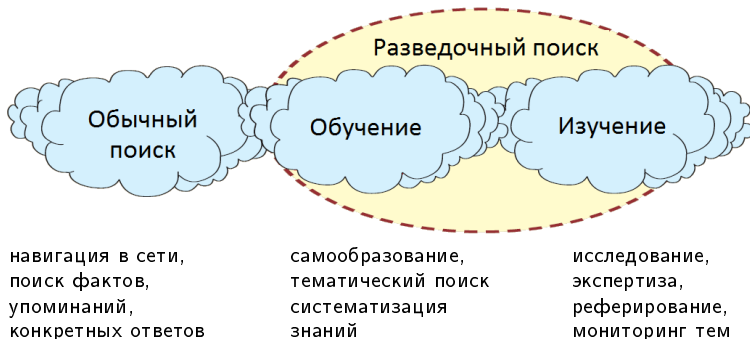
**Лемма 3.** Формулы М-шага:

$$\phi_{wt} = \operatorname{norm}_{w \in W} \left( \sum_{d \in D} n_{dw} p_{tdw} + \sum_{d \in D} Q_{tdw} p_{tdw} + \phi_{wt} \frac{\partial \tilde{R}}{\partial \phi_{wt}} \right);$$

$$\theta_{td} = \operatorname{norm}_{t \in T} \left( \sum_{w \in D} n_{dw} p_{tdw} + \sum_{w \in D} Q_{tdw} p_{tdw} + \theta_{td} \frac{\partial \tilde{R}}{\partial \theta_{td}} \right).$$

## Концепция разведочного поиска (Exploratory Search)

- пользователь может не знать ключевых терминов,
- запросом может быть текст произвольной длины,
- информационной потребностью — систематизация знаний



Gary Marchionini. Exploratory Search: from finding to understanding. 2006.

## Две коллекции новостей про технологии

### Habrhabr.ru

175 143 статей на русском  
10 552 слов (униграмм)  
742 000 биграмм  
524 авторов статей  
10 000 авторов комментариев  
2546 тегов  
123 хаба (категории)

### TechCrunch.com

759 324 статей на английском  
11 523 слов (униграмм)  
1.2 млн. биграмм  
605 авторов  
184 категорий



## Методика оценивания качества разведочного поиска

### Поисковый запрос

набор ключевых слов или фрагментов текста, около одной страницы A4

### Поисковая выдача

документы  $d$  с распределением  $p(t|d)$ , близким к распределению  $p(t|q)$  запроса

### Два задания асессорам

- найти как можно больше статей, пользуясь любыми средствами поиска (и засечь время)
- оценить релевантность поисковой выдачи на том же запросе

#### Поиск MapReduce

**Поиск MapReduce** – программа поиска (**Байесовый**) вычислений распределенных вычислений для больших объемов данных и разная парадигма поиска, представляющая собой набор Java-классов и исполняемых утилит для создания и обработки данных на параллельной обработке.

**Основные компоненты** **Поиск MapReduce** имеют следующие характеристики:

- обработка вычислений больших объемов данных;
- масштабируемость;
- автоматическое распределение заданий;
- работа на выделенных оборудовании;
- автоматическая обработка отказов вычислений заданий.

**Поиск** – популярная программная платформа (**Яндекс.Байесовый**) построена распределенных приложений для массово-параллельной обработки (**разное разбит** **русский** **МР**) данных.

**Поиск** включает в себе следующие компоненты:

1. **MRFS** – распределенная файловая система;
2. **Поиск MapReduce** – программная платформа (**Байесовый**) вычислений распределенных вычислений для больших объемов данных и разная парадигма поиска;

**Клиенты, приложения и архитектура** **Поиск MapReduce** и структура **MRFS**, стали прототипом ряда других систем в области вычислений, в том числе и в отношении точки отказа. Это, в конечном итоге, определило направление платформ **Поиск** и сейчас. К последним можно отнести:

Ограничение масштабируемости кластера **Поиск** – это вычислительный узел, «ИК параллельных заданий.

Сильная связность **Фреймворк** распределенных вычислений и клиентских вычислений реализован распределенной архитектурой. Как следствие:

Существенная поддержка алгоритмической программы вычислений распределенных вычислений в **Поиск v1.0** поддерживается только модель вычислений **параллельно**.

Модель вычислений, точки отказа и как следствие, масштабируемость вычислений в среде с высоким требованием к надежности.

Проблема **взаимосвязи** совместности требования по единичному вычислению вычисления всех вычислительных узлов кластера при обилии платформ **Поиск** (установка новой версии или пакета обновлений).

Пример запроса для разведочного поиска



## Пример: фрагмент запроса «Система IBM Watson»

IBM Watson — суперкомпьютер фирмы IBM, оснащённый вопросно-ответной системой искусственного интеллекта, созданный группой исследователей под руководством Дэвида Феруччи. Его создание — часть проекта DeepQA. Основная задача Уотсона — понимать вопросы, сформулированные на естественном языке, и находить на них ответы в базе данных. Назван в честь основателя IBM Томаса Уотсона.

IBM Watson представляет собой когнитивную систему, которая способна понимать, делать выводы и обучаться. Она также позволяет преобразовывать целые отрасли, различные направления науки и техники. Например, предсказывать появление эпидемий или возникновения очагов природных катастроф в различных регионах, вести мониторинг состояния атмосферы больших городов, оптимизировать бизнес-процессы, узнавать, какие товары будут в тренде в ближайшее время.

... ..

**Релевантные тексты:** примеры сервисов и приложений, основа которых — когнитивная платформа IBM Watson, используемые в IBM Watson технологии, вопрос-ответные системы, сопоставление IBM Watson с Wolfram-Alpha.

**Нерелевантные тексты:** общие вопросы искусственного интеллекта, другие коммерческие решения на рынке бизнес-аналитики.

## Тематика запросов разведочного поиска

Примеры заголовков разведочных запросов к Хабру  
(объём каждого запроса — около одной страницы А4):

Алгоритмы раскраски графов	Система IBM Watson
Рекомендательная система Netflix	3D-принтеры
Методики быстрого набора текста	CERN-кластер
Космические проекты Илона Маска	АВ-тестирование
Технологии Hadoop MapReduce	Облачные сервисы
Беспилотный автомобиль Google car	Контекстная реклама
Криптосистемы с открытым ключом	Марсоход Curiosity
Обзор платформ онлайн-курсов	Видеокарты NVIDIA
Data Science Meetups в Москве	Распознавание образов
Образовательные проекты mail.ru	Сервисы Google scholar
Межпланетная станция New horizons	MIT MediaLab Research
Языковая модель word2vec	Платформа Microsoft Azure

## Разведочный тематический поиск

$q = (w_1, \dots, w_{n_q})$  — текст запроса произвольной длины  $n_q$

$\theta_{tq} = p(t|q)$  — тематический профиль запроса  $q$

$\theta_{td} = p(t|d)$  — тематические профили документов  $d \in D$

Косинусная мера близости документа  $d$  и запроса  $q$ :

$$\text{sim}(q, d) = \frac{\sum_t \theta_{tq} \theta_{td}}{(\sum_t \theta_{tq}^2)^{1/2} (\sum_t \theta_{td}^2)^{1/2}}.$$

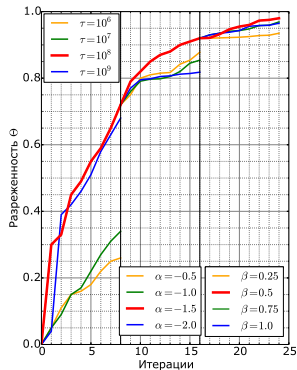
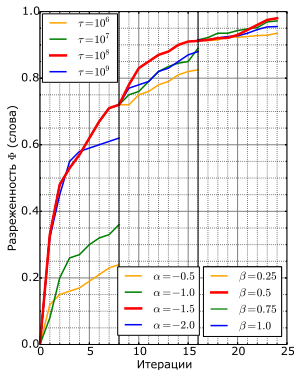
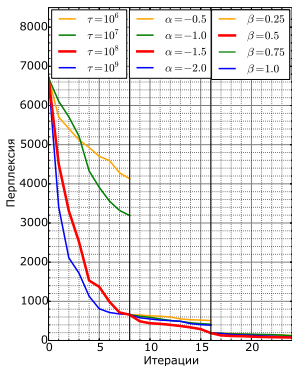
Ранжируем документы коллекции  $d \in D$  по убыванию  $\text{sim}(q, d)$

Выдача тематического поиска —  $k$  первых документов.

Реализация: *инвертированный индекс* для быстрого поиска документов  $d$  по каждой из тем  $t$  запроса

## Последовательный подбор коэффициентов регуляризации

- декоррелирование распределений терминов в темах ( $\tau$ ),
- разреживание распределений тем в документах ( $\alpha$ ),
- сглаживание распределений терминов в темах ( $\beta$ ).



## Оценки качества поиска

Precision — доля релевантных среди найденных

Recall — доля найденных среди релевантных

$$P = \frac{TP}{TP + FP} \text{ — точность (precision)}$$

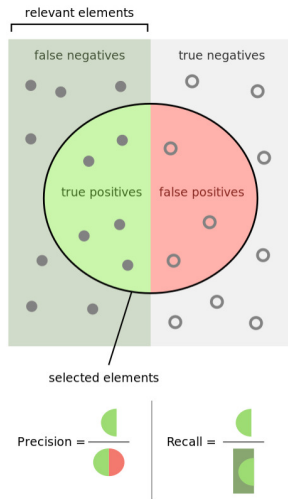
$$R = \frac{TP}{TP + FN} \text{ — полнота, (recall)}$$

$$F_1 = \frac{P + R}{2PR} \text{ — F1-мера}$$

TP (true positive) — найденные релевантные

FP (false positive) — найденные нерелевантные

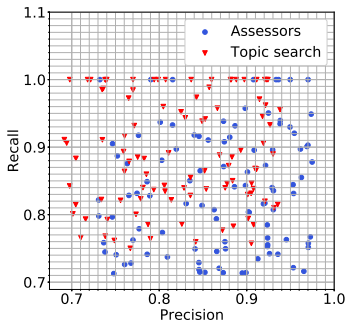
FN (false negative) — не найденные релевантные



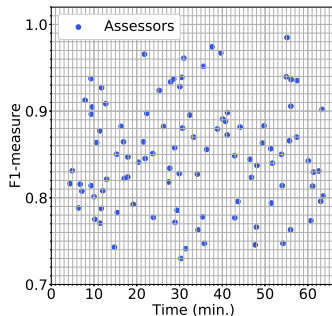
## Результаты измерения точности и полноты по запросам

100 запросов, 3 ассессора на запрос

точность и полнота поиска



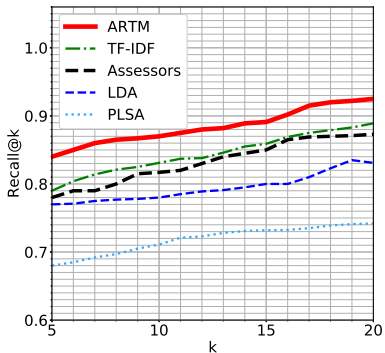
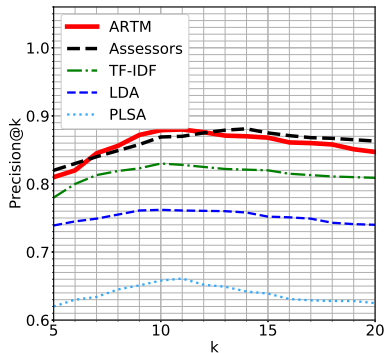
время и  $F_1$ -мера (ассессоры)



- среднее время обработки запроса ассессором — 30 минут
- полнота чуть лучше, точность чуть хуже, чем у ассессоров

## Сравнение с ассессорами по качеству поиска

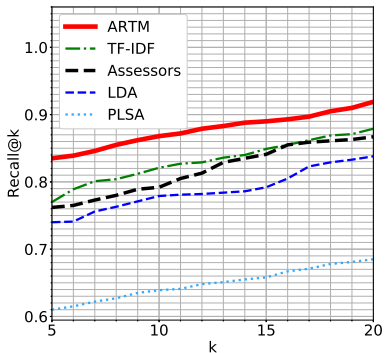
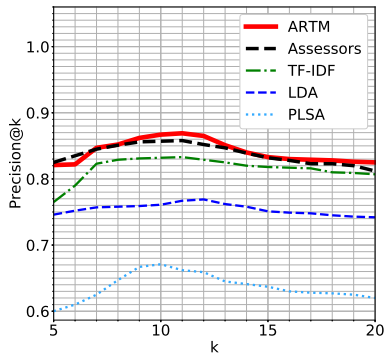
Точность и полнота по первым  $k$  позициям поисковой выдачи (коллекция Habrhabr.ru)



A. Ianina, K. Vorontsov. Multi-objective topic modeling for exploratory search in tech news. AINL, 2017.

## Сравнение с ассессорами по качеству поиска

Точность и полнота по первым  $k$  позициям поисковой выдачи (коллекция TechCrunch.com)



A. Ianina, K. Vorontsov. Multi-objective topic modeling for exploratory search in tech news. AINL, 2017.













## Влияние комбинаций регуляризаторов на качество поиска

Декоррелирование, Θ-разреживание, Φ-сглаживание

	Habrahabr				TechCrunch			
	$R = 0$	Д	ДΘ	ДΘΦ	$R = 0$	Д	ДΘ	ДΘΦ
Prec@5	0.628	0.748	0.771	<b>0.810</b>	0.652	0.775	0.779	<b>0.819</b>
Prec@10	0.653	0.776	0.812	<b>0.879</b>	0.679	0.787	0.819	<b>0.867</b>
Prec@15	0.642	0.765	0.792	<b>0.868</b>	0.669	0.773	0.798	<b>0.833</b>
Prec@20	0.643	0.759	0.783	<b>0.847</b>	0.673	0.777	0.792	<b>0.825</b>
Recall@5	0.692	0.784	0.805	<b>0.840</b>	0.673	0.812	0.812	<b>0.835</b>
Recall@10	0.714	0.814	0.834	<b>0.870</b>	0.685	0.821	0.845	<b>0.868</b>
Recall@15	0.725	0.835	0.867	<b>0.891</b>	0.712	0.859	0.869	<b>0.890</b>
Recall@20	0.735	0.862	0.891	<b>0.925</b>	0.723	0.882	0.895	<b>0.919</b>

- комбинирование регуляризаторов улучшает качество поиска
- хотя исходно все регуляризаторы нацелены на улучшение интерпретируемости тем и не оптимизируют поиск явно

- Тематическое моделирование — средство семантического поиска и систематизации текстовой информации
- ARTM — многокритериальная регуляризация
- BigARTM — «ЛЕГО-конструктор» тематических моделей
- Регуляризаторы  $n$ -грамм, совстречаемости и сегментации позволяют обходить гипотезу «мешка слов»
- Для автоматического выделения терминов синтаксический анализ можно заменить тематическим
- Тематические модели строят разреженные интерпретируемые векторные представления слов и решают задачу семантической близости не хуже word2vec

-  *K.V.Vorontsov*. Обзор вероятностных тематических моделей. 2017. – **NEW!**  
<http://www.MachineLearning.ru/wiki/images/d/d5/Voron17survey-artm.pdf>
-  *K.V.Vorontsov*. Аддитивная регуляризация тематических моделей коллекций текстовых документов. Доклады РАН. 2014.
-  *K.Vorontsov, A.Potapenko*. Additive regularization of topic models. Machine Learning, 2015.
-  *K.Vorontsov, O.Frei, M.Apishev, P.Romov, M.Suvorova, A.Yanina*. Non-bayesian additive regularization for multimodal topic modeling of large collections. 2015.
-  *K.Vorontsov, A.Potapenko, A.Plavin*. Additive regularization of topic models for topic selection and sparse factorization. SLDS 2015.
-  *O.Frei, M.Apishev*. Parallel non-blocking deterministic algorithm for online topic modeling. AIST 2016.
-  *M.Apishev, S.Koltcov, O.Koltsova, S.Nikolenko, K.Vorontsov*. Additive regularization for topic modeling in sociological studies of user-generated text content. MICAI 2016.
-  *N.Chirkova, K.Vorontsov*. Additive regularization for hierarchical multimodal topic modeling. JMLDA, 2016.
-  *A.Ianina, K.Vorontsov*. Multi-objective topic modeling for exploratory search in tech news. AINL, 2017.
-  *A.Potapenko, A.Popov, K.Vorontsov*. Interpretable probabilistic embeddings: bridging the gap between topic models and neural networks. AINL, 2017.