Аддитивно регуляризованные тематические модели и разведочный поиск знаний в сети

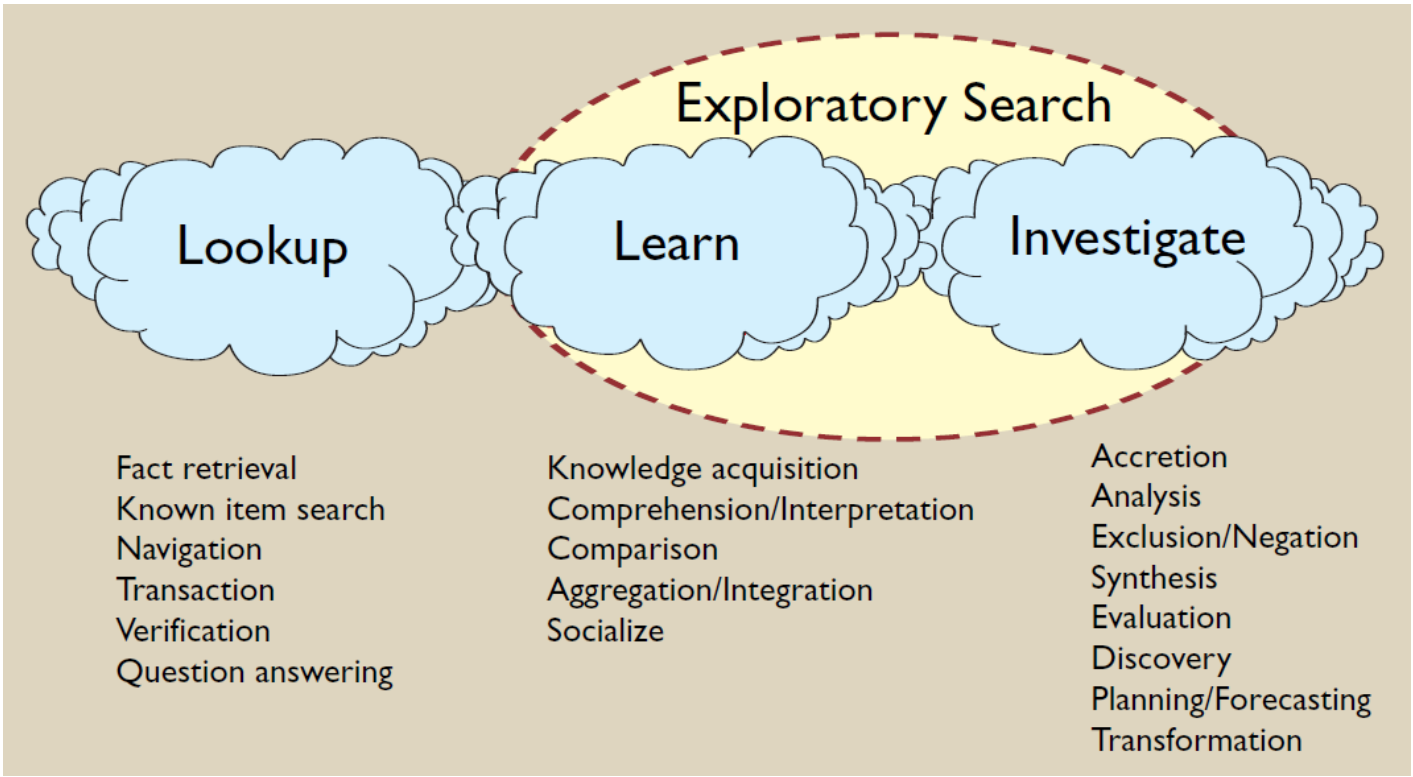
Воронцов Константин Вячеславович
ВЦ РАН • МФТИ • МГУ • ВШЭ • Яндекс • FORECSYS

• Традиционная молодёжная летняя школа •
19 июня 2015

Contents

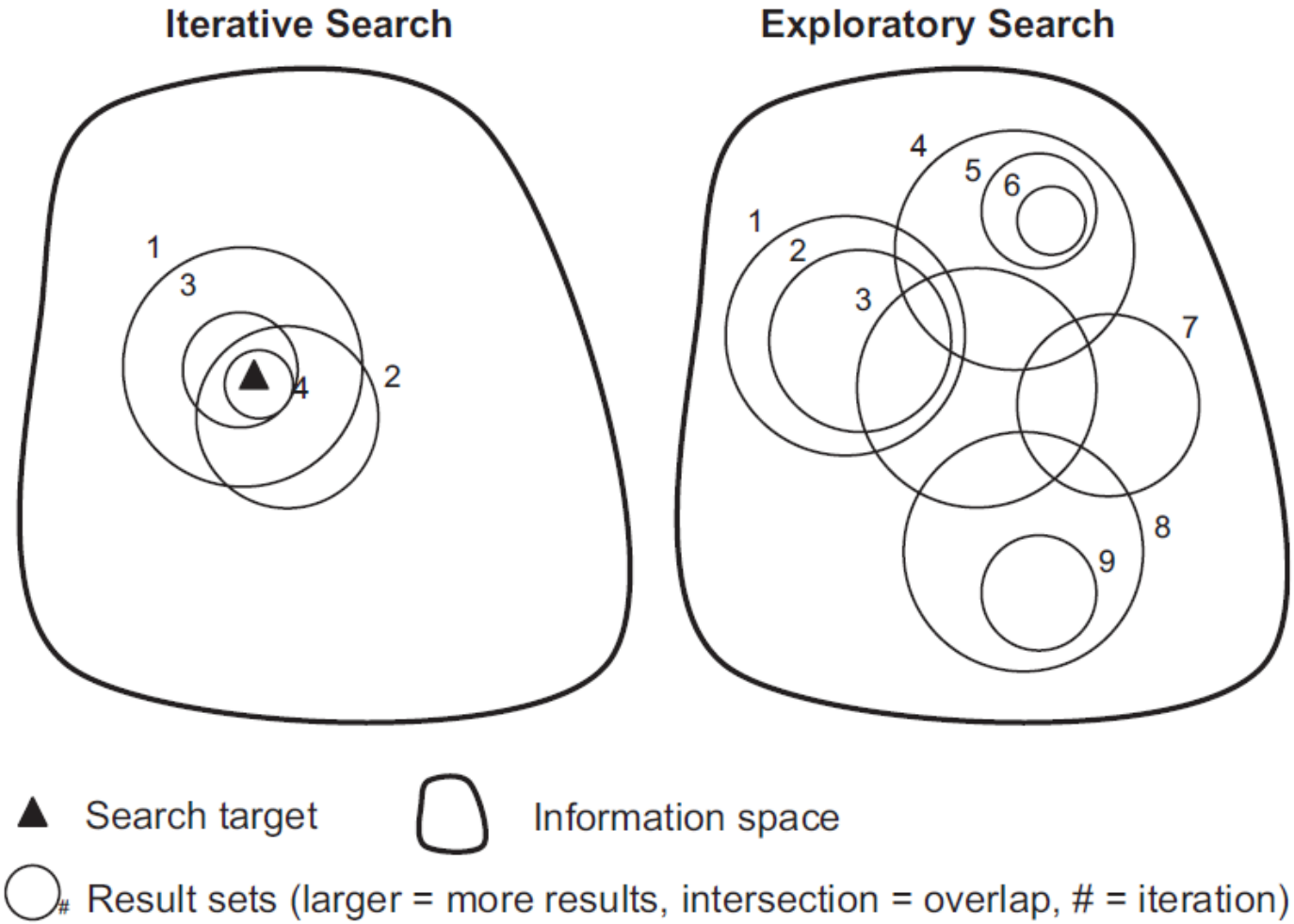
Разведочный поиск как инструмент самообразования

- пользователь может не знать ключевых терминов
- пользователя может интересовать множество ответов



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

От поиска “query-browse-refine” к разведочному поиску



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

Возможный сценарий разведочного поиска

Поисковый запрос:

- документ любой длины или даже коллекция документов

Цели поиска:

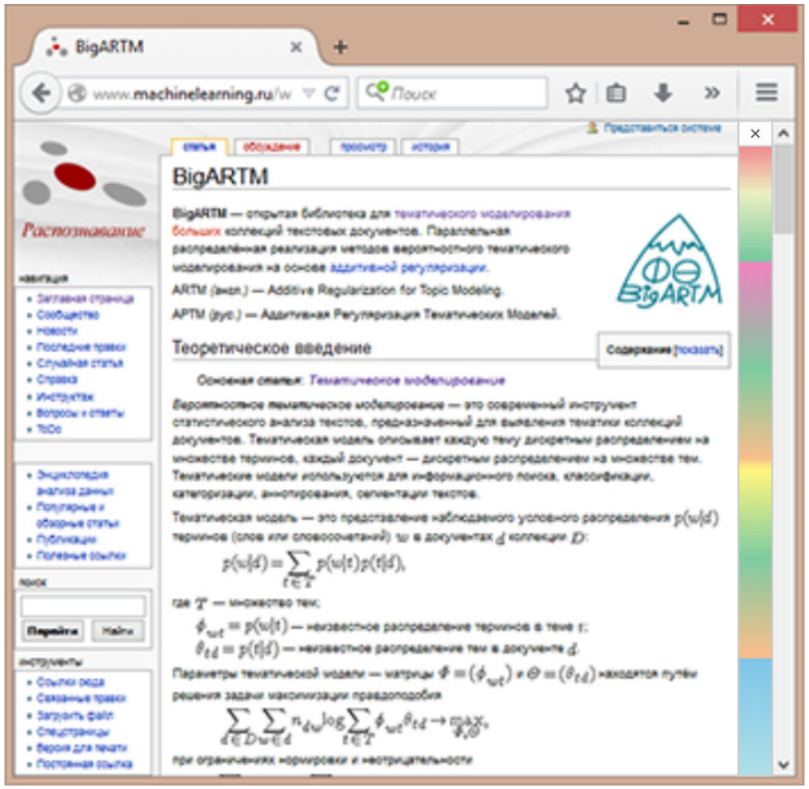
- к каким темам относится мой запрос?
- что ещё известно по этим темам?
- какова тематическая структура этой предметной области?
- что ещё есть понятного, обзорного, важного, свежего?

Сценарий поиска:

1. имея любой текст под рукой, в любом приложении,
2. хотим получить картину содержащихся в нём тем-подтем,
3. и «дорожную карту» предметной области в целом

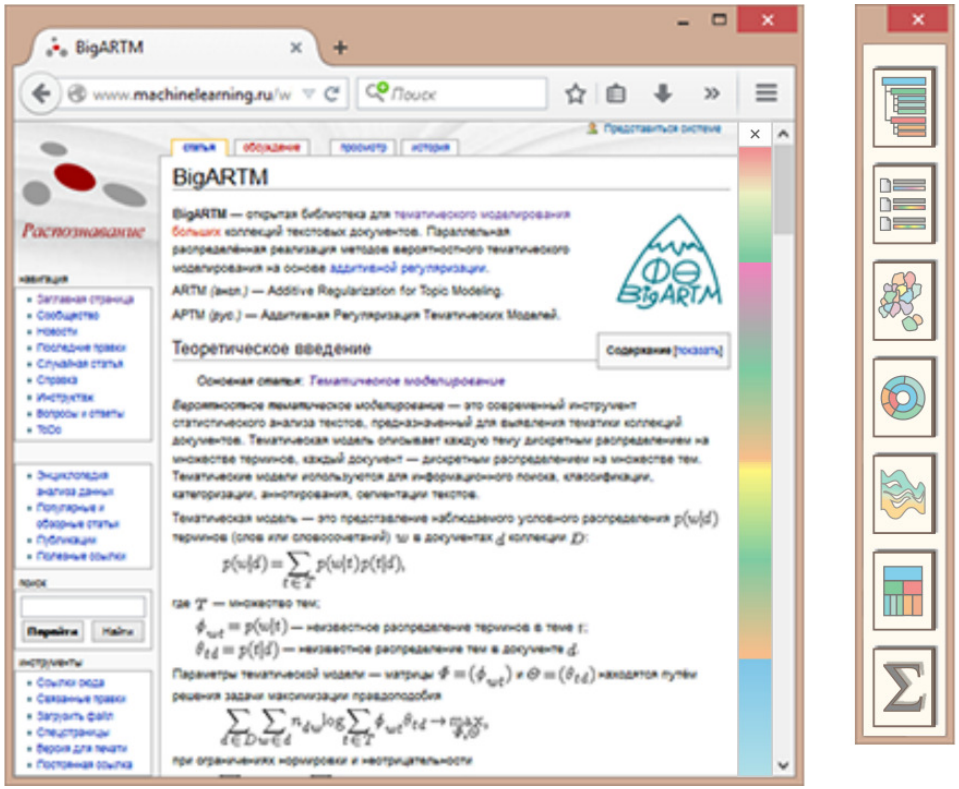
Разведочный поиск: прототип интерфейса

Радужная полоса напоминает, что знания всегда под рукой



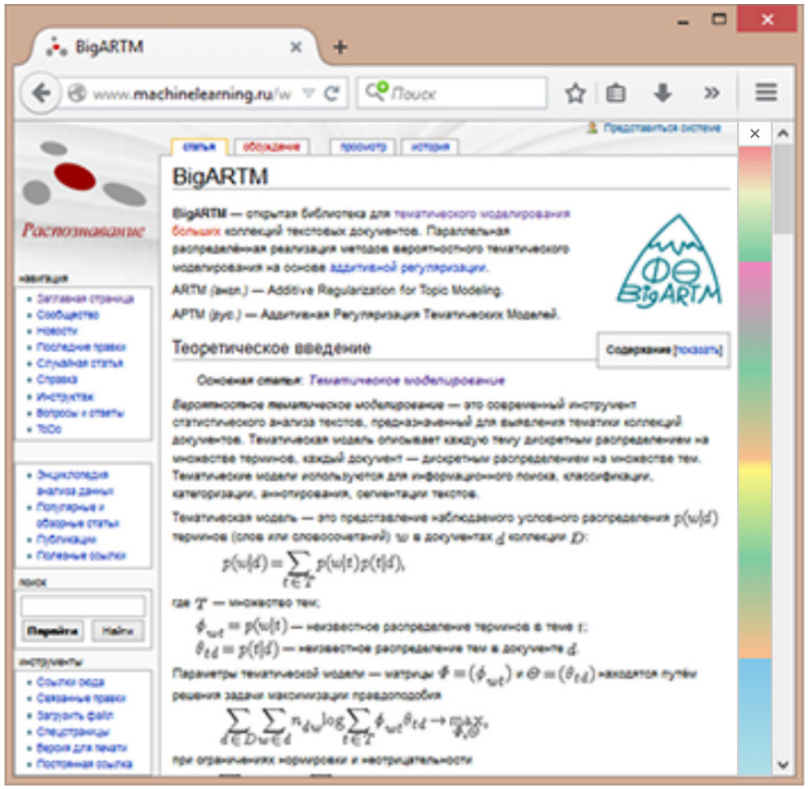
Разведочный поиск: прототип интерфейса

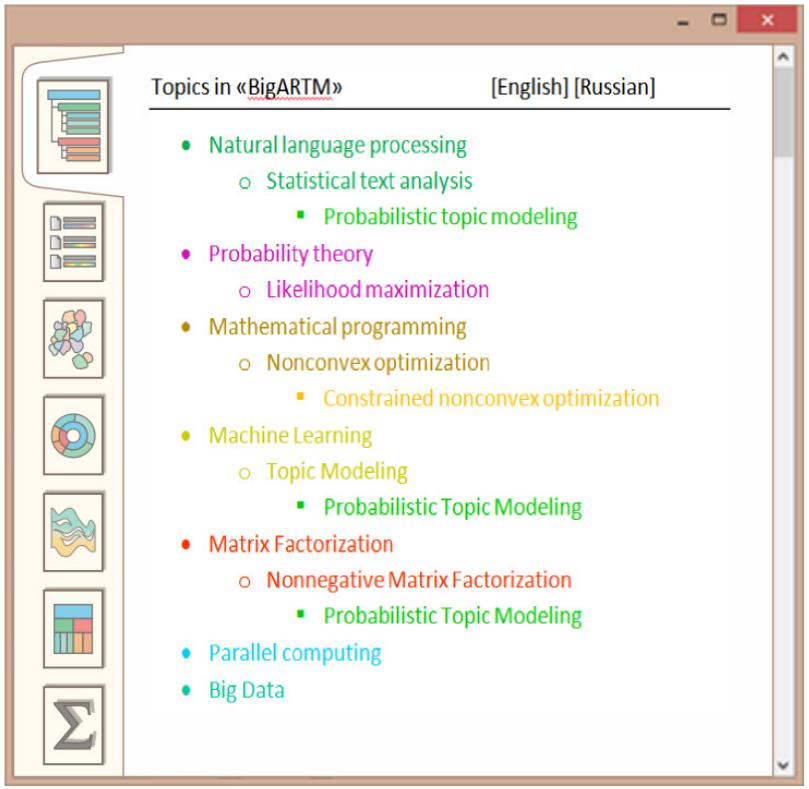
Клик по радужной полосе — тематический поисковый запрос



Разведочный поиск: прототип интерфейса

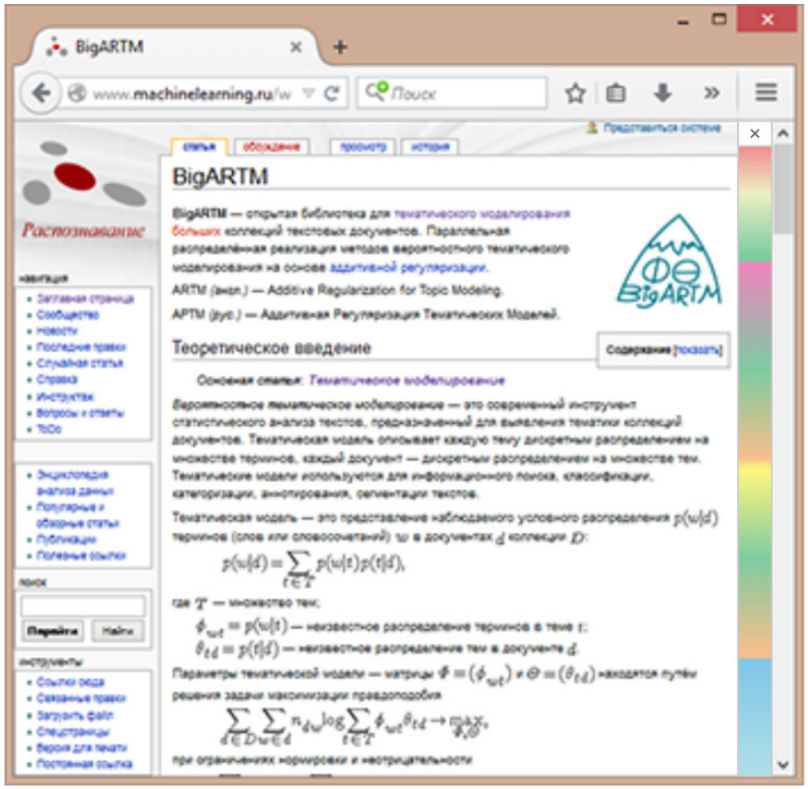
Темы-подтемы выбранного фрагмента текста

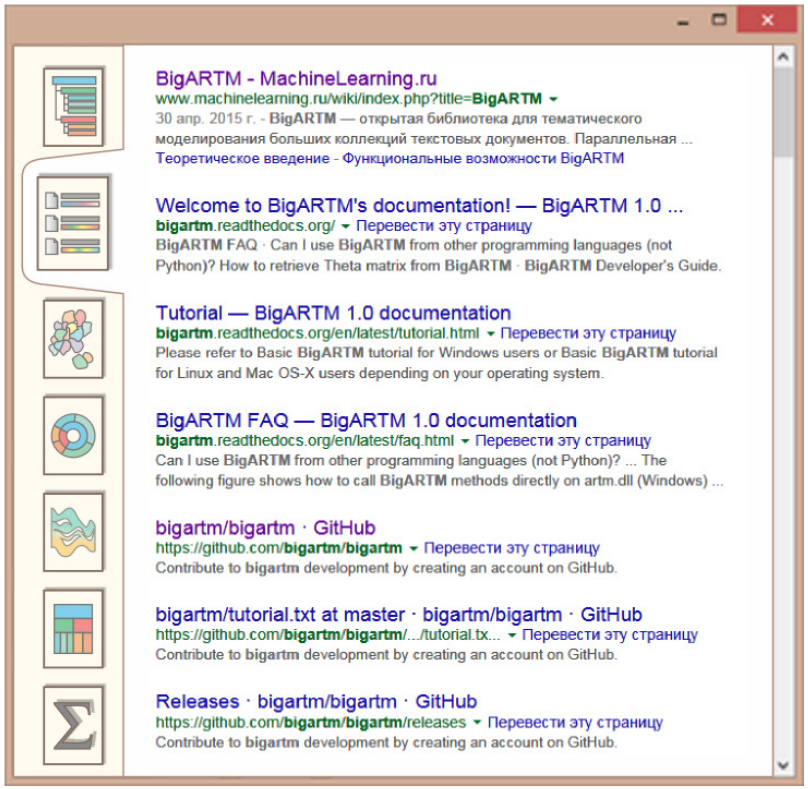




Разведочный поиск: прототип интерфейса

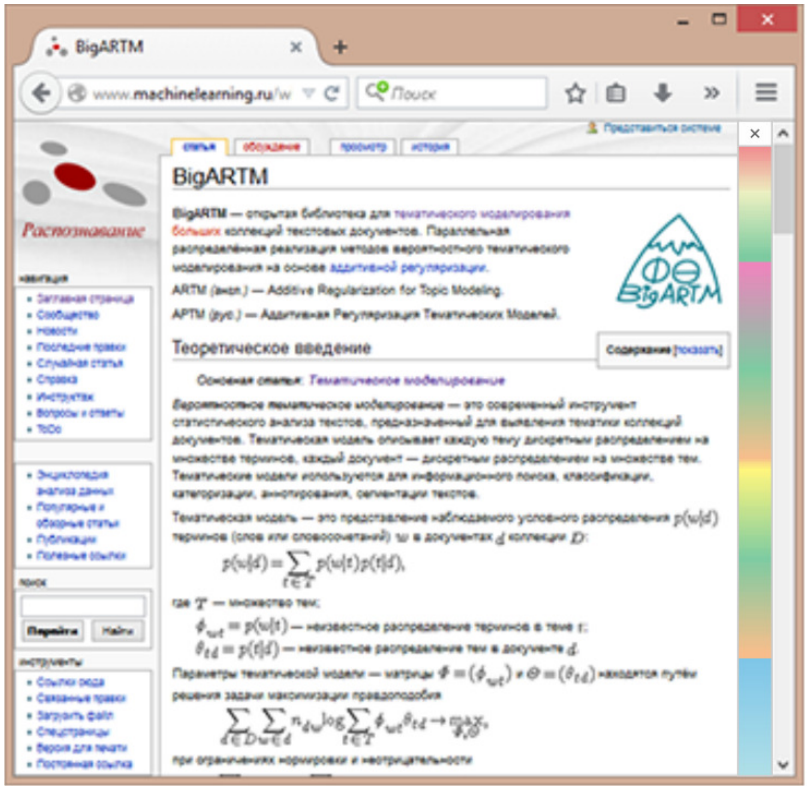
Документы и иные объекты, ранжированные по релевантности

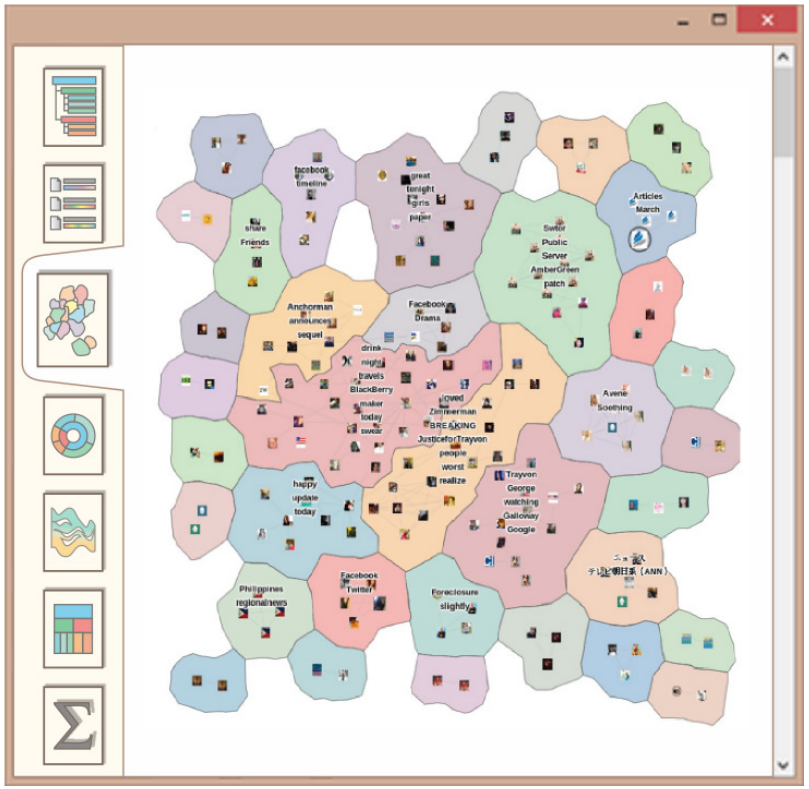




Разведочный поиск: прототип интерфейса

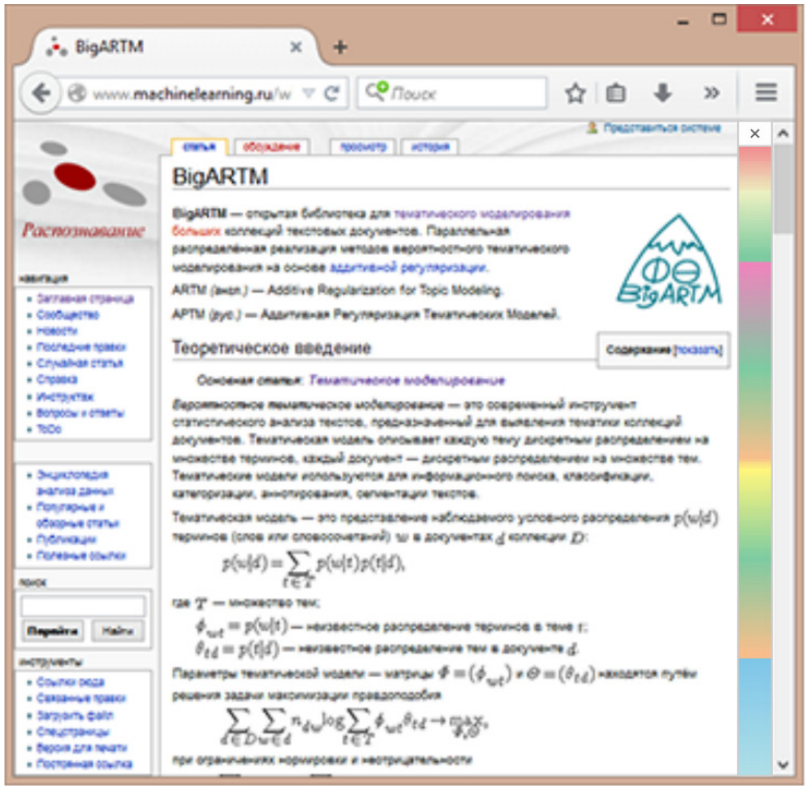
Дорожная карта: кластеризация релевантных документов

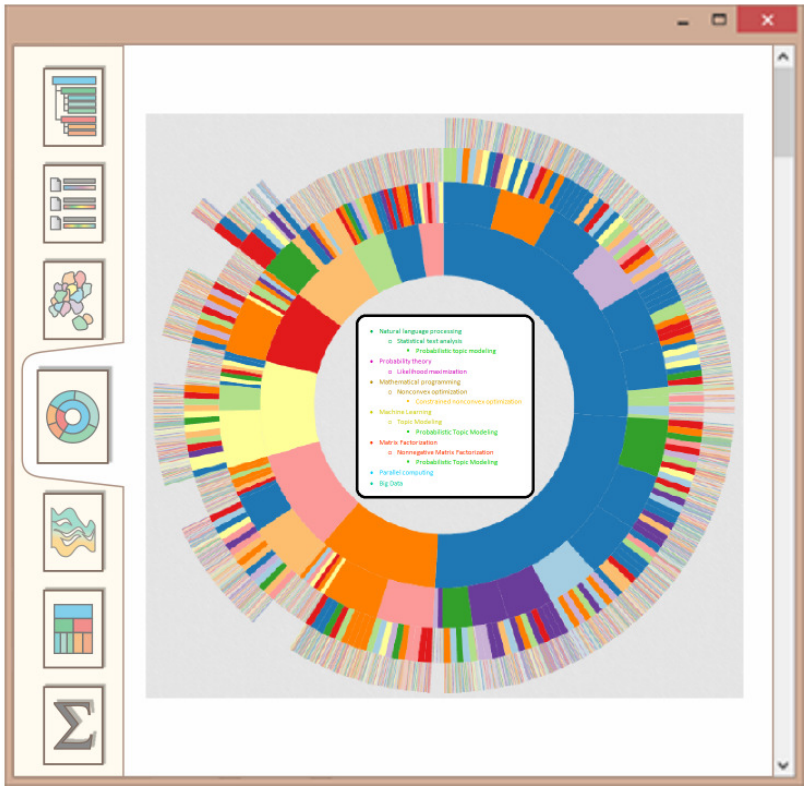




Разведочный поиск: прототип интерфейса

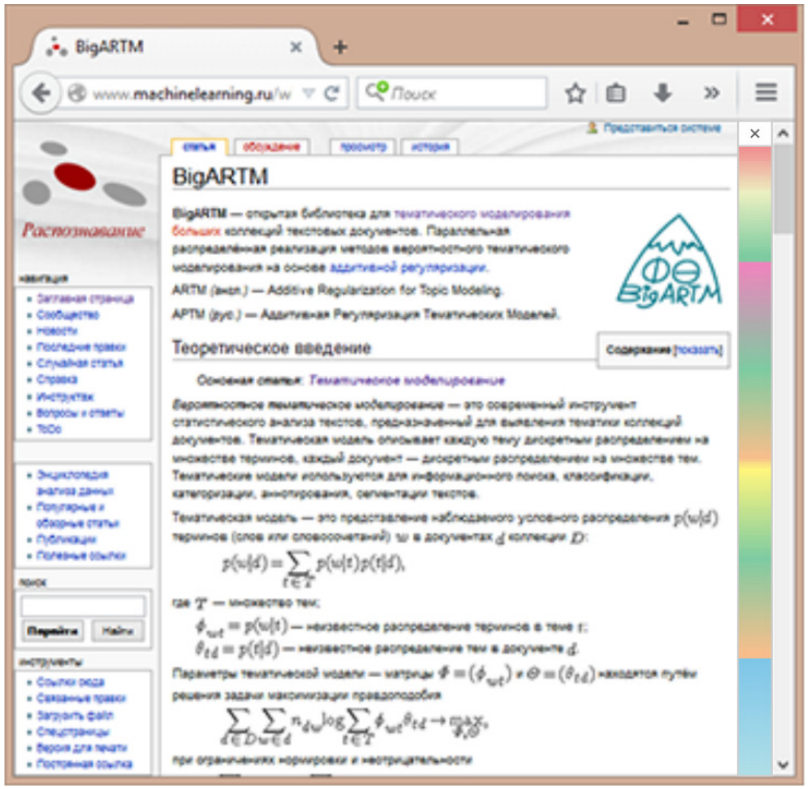
Тематическая иерархия: структура предметной области

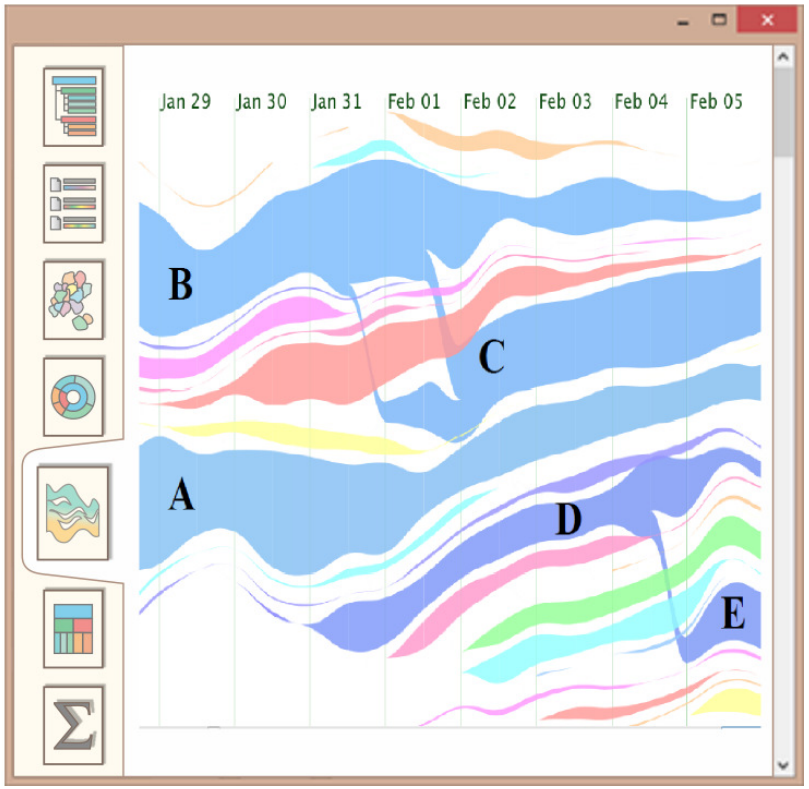




Разведочный поиск: прототип интерфейса

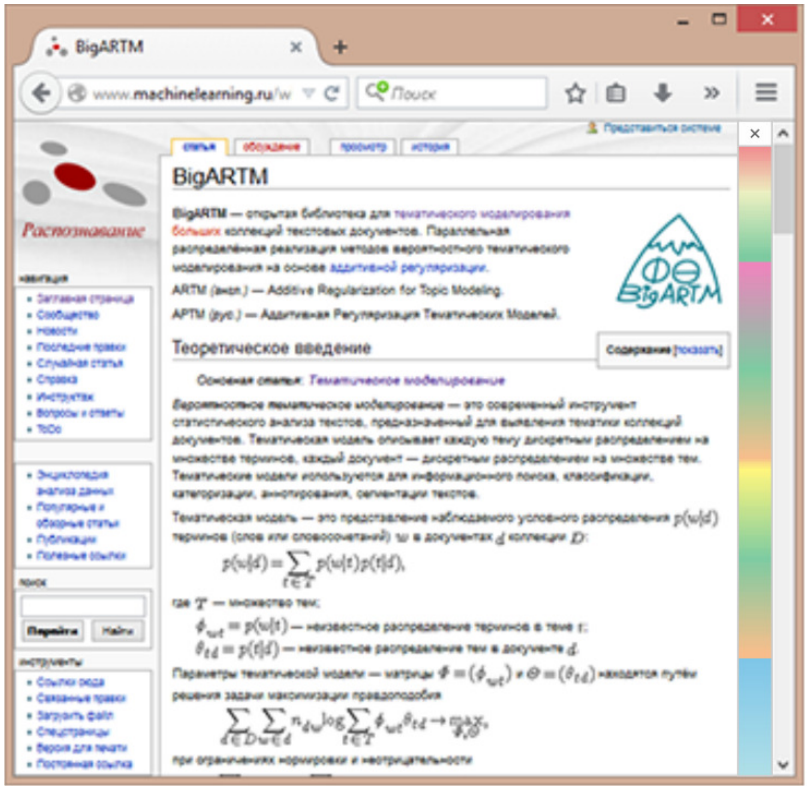
Динамика тем: эволюция предметной области

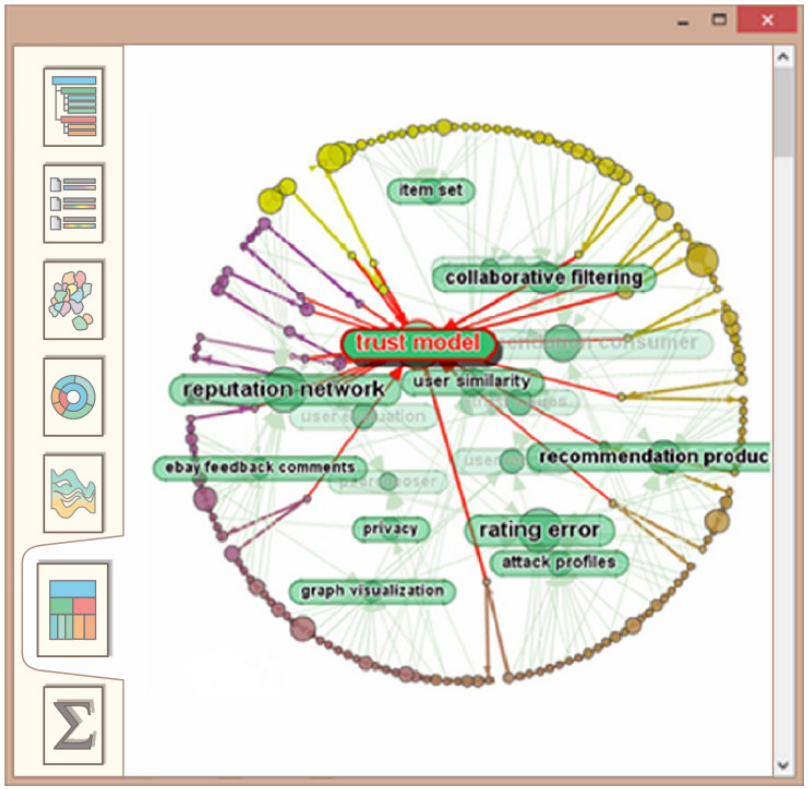




Разведочный поиск: прототип интерфейса

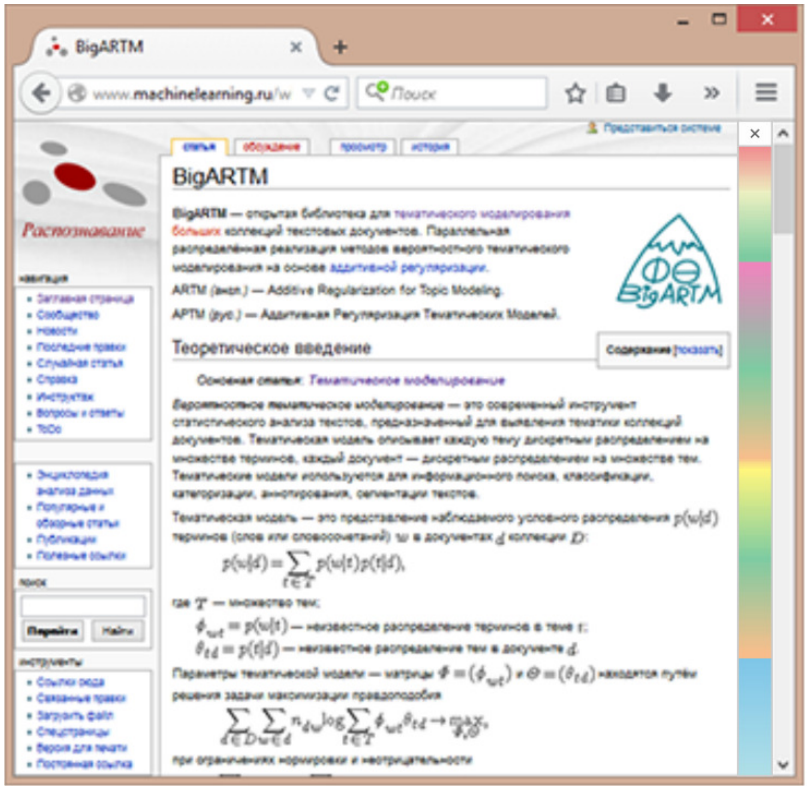
Тематическая сегментация документа запроса

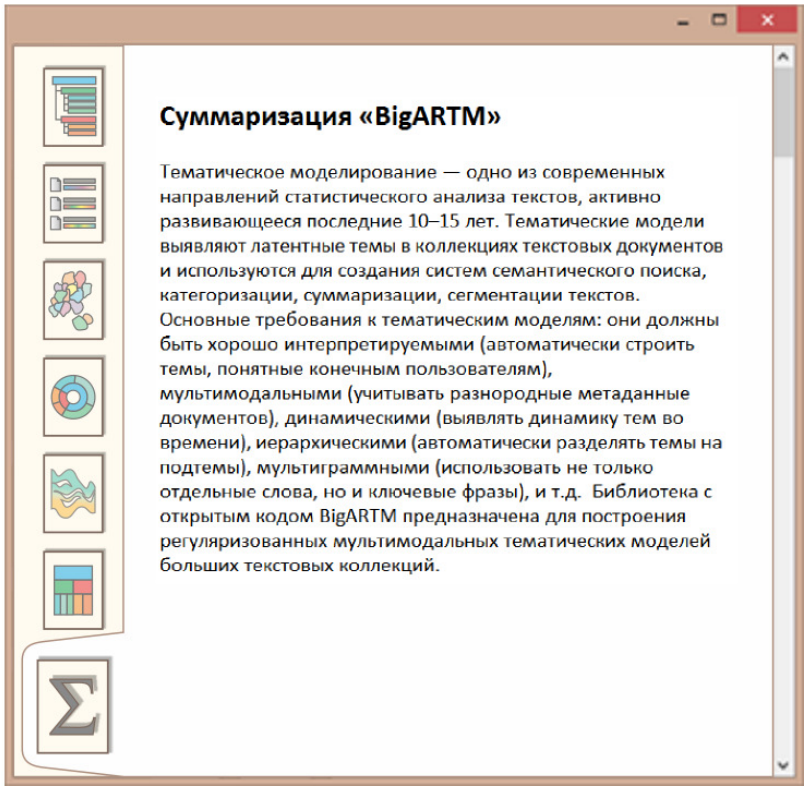




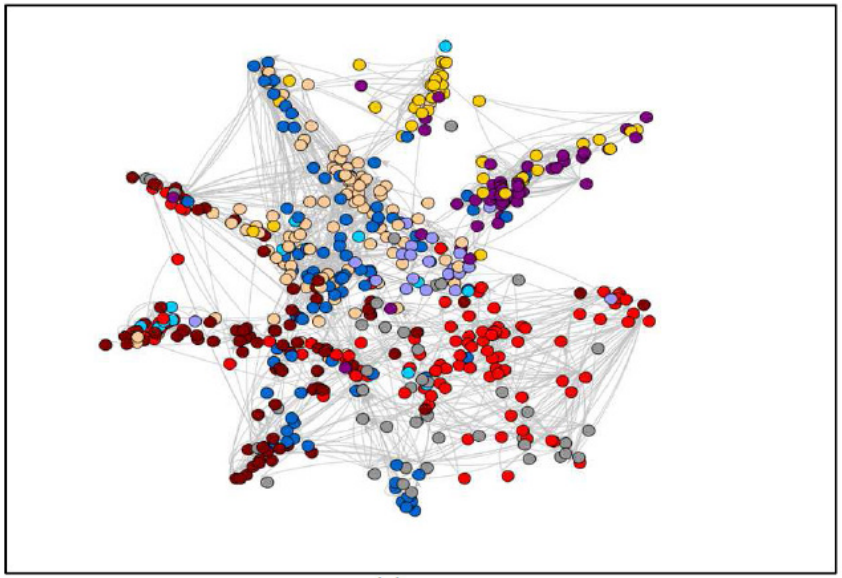
Разведочный поиск: прототип интерфейса

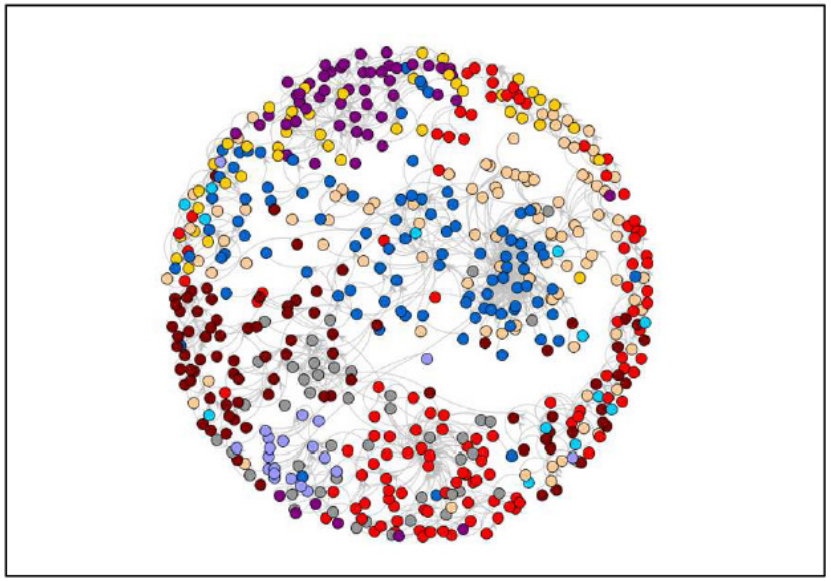
Суммаризация документа запроса





Дорожная карта: кластеризация релевантных документов

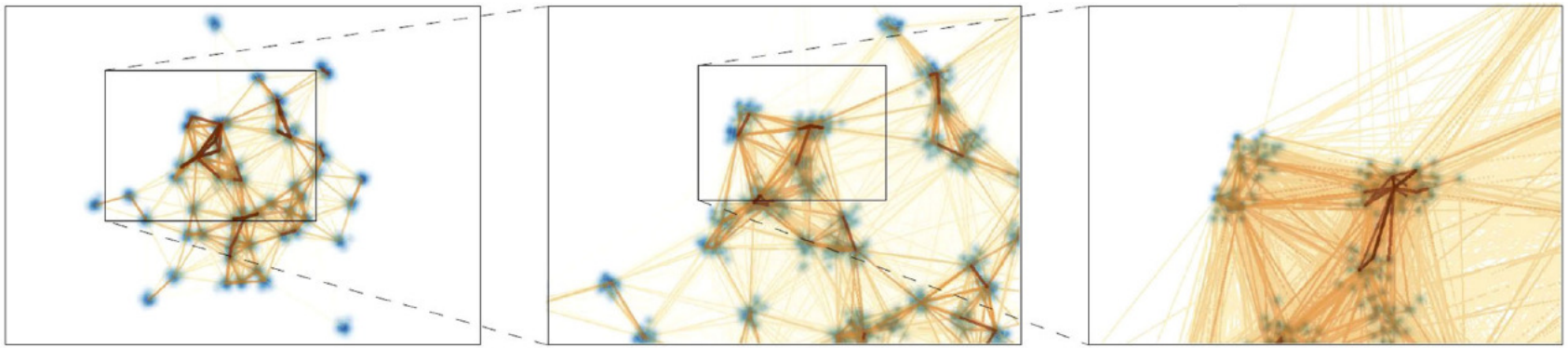




- Точки — это документы (или их фрагменты)
- Кластеры — это группы тематически схожих документов
- Форму облака точек можно удобно настраивать

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

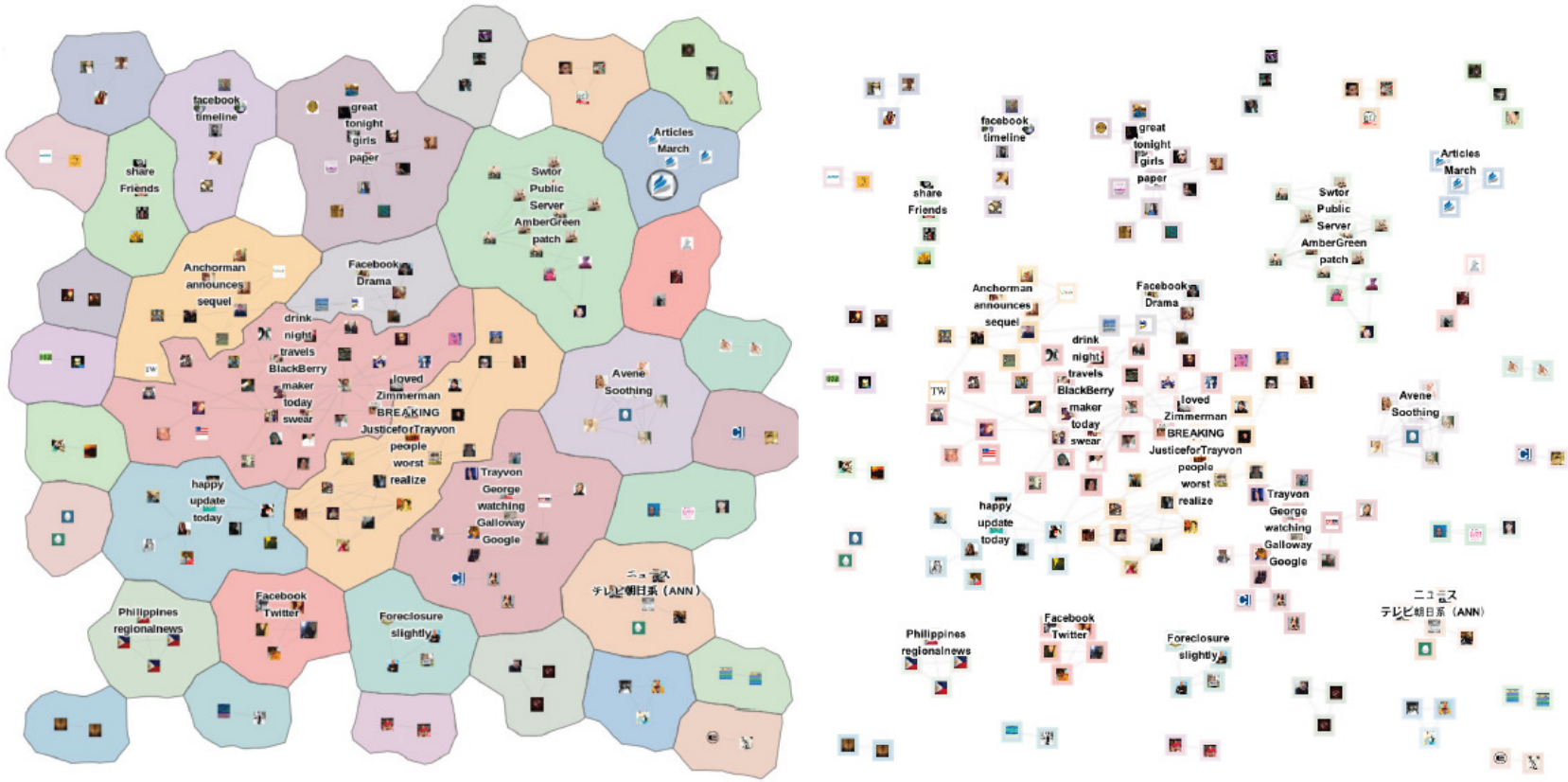
Дорожная карта: кластеризация релевантных документов



- Кластеры
  - кластеров
    - кластеров
      - кластеров...

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

Дорожная карта: кластеризация релевантных документов



«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.

http://textvis.lnu.se

Интерактивный обзор 170 средств визуализации текстов

Технологические элементы разведочного поиска

1. Интернет-краулинг ..... имеются готовые решения
2. Фильтрация контента ..... имеются готовые решения
3. Тематическое моделирование ..... **математика здесь**
4. Инвертированный индекс ..... имеются готовые решения
5. Ранжирование ..... имеются готовые решения
6. Визуализация ..... имеются готовые решения

Тематическая модель для разведочного поиска должна быть...

1. **Интерпретируемая**: каждая тема понятна людям
2. **Мультиграммная**: термины-словосочетания неразрывны
3. **Мультиязычная**: кросс-языковой и много-языковой поиск
4. **Мультимодальная**: авторы, связи, тэги, пользователи,...
5. **Динамическая**: развитие тем во времени
6. **Иерархическая**: настраиваемая гранулярность тем
7. **Сегментирующая**: границы тем внутри документа
8. **Обучаемая**: учёт экспертных оценок

Недостатки тематических моделей для разведочного поиска

1. Плохо развиты техники комбинирования моделей

И пути их устранения:

1. ARTM: Аддитивная Регуляризация Тематических Моделей

Недостатки тематических моделей для разведочного поиска

1. Плохо развиты техники комбинирования моделей
2. Интерпретируемость пока не достигается автоматически

И пути их устранения:

1. ARTM: Аддитивная Регуляризация Тематических Моделей
2. • автоматическое выделение терминов n-грамм
   • использование внешних лингвистических ресурсов
   • выделение лексических ядер тем (разреженность, различность, согласованность)

Недостатки тематических моделей для разведочного поиска

1. Плохо развиты техники комбинирования моделей
2. Интерпретируемость пока не достигается автоматически
3. Не достаточно используются лингвистические знания

И пути их устранения:

1. ARTM: Аддитивная Регуляризация Тематических Моделей
2. • автоматическое выделение терминов n-грамм
   • использование внешних лингвистических ресурсов
   • выделение лексических ядер тем (разреженность, различность, согласованность)
3. Лингвистическая регуляризация тематических моделей (sentence TM, syntactic TM, segmentation TM,. . .)

Что такое «тема»?

- *Тема* — специальная терминология предметной области.
- *Тема* — набор терминов (слов или словосочетаний), совместно часто встречающихся в документах.

Более формально,

- *тема* — условное распределение на множестве терминов, $p(w|t)$ — вероятность термина w в теме t;
- *тематический профиль* документа — условное распределение $p(t|d)$ — вероятность темы t в документе d.

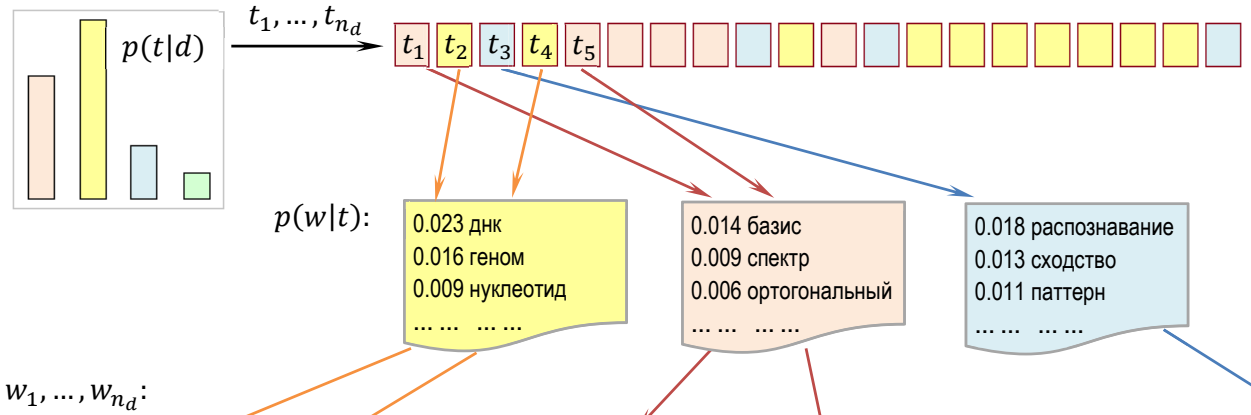
Когда автор писал термин w в документ d, он думал о теме t, и мы хотели бы догадаться, о какой именно

Тематическая модель выявляет латентные темы по наблюдаемым распределениям слов $p(w|d)$ в документах.

Прямая задача — порождение коллекции по $p(w|t)$ и $p(t|d)$

Вероятностная тематическая модель коллекции документов D:

$$p(w|d) = \sum_t p(w|t)p(t|d), \quad d \in D$$



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).

Обратная задача — восстановление $p(w|t)$ и $p(t|d)$ по коллекции

Дано: W — словарь терминов
D — коллекция текстовых документов $d = \{w_1 \dots w_{n_d}\}$
n_{dw} = сколько раз термин w встречается в документе d

Найти параметры модели $p(w|d) = \sum_t \phi_{wt}\theta_{td}$:

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

Задача максимизации логарифма правдоподобия:

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

при ограничениях нормировки и неотрицательности

$$\phi_{wt} \geqslant 0; \quad \sum_w \phi_{wt} = 1; \qquad \theta_{td} \geqslant 0; \quad \sum_t \theta_{td} = 1$$

PLSA — Probabilistic Latent Semantic Analysis [Hofmann, 1999]

Теорема

Точка максимума $\mathscr{L}(\Phi, \Theta)$ удовлетворяет системе уравнений со вспомогательными переменными $p_{tdw} \equiv p(t|d,w)$, n_{wt}, n_{td}:

$$\text{E-шаг:} \quad p_{tdw} = \frac{\phi_{wt}\theta_{td}}{\sum_{t'} \phi_{wt'}\theta_{t'd}}$$

$$\text{M-шаг:} \quad \begin{cases} \phi_{wt} = \dfrac{n_{wt}}{n_t}; \quad n_{wt} = \sum\limits_{d \in D} n_{dw} p_{tdw}; \quad n_t = \sum\limits_{w \in W} n_{wt} \\ \theta_{td} = \dfrac{n_{td}}{n_d}; \quad n_{td} = \sum\limits_{w \in d} n_{dw} p_{tdw}; \quad n_d = \sum\limits_{t \in T} n_{td} \end{cases}$$

EM-алгоритм — чередование E- и M-шага до сходимости,
т. е. решение системы уравнений методом простых итераций.

✓ *Идея на будущее: можно использовать и другие методы!*

EM-алгоритм. Элементарная интерпретация

EM-алгоритм — это чередование E и M шагов до сходимости.

E-шаг: условные вероятности тем $p(t|d,w)$ для всех t, d, w вычисляются через ϕ_{wt}, θ_{td} по формуле Байеса:

$$p(t|d,w) = \frac{p(w,t|d)}{p(w|d)} = \frac{p(w|t)p(t|d)}{p(w|d)} = \frac{\phi_{wt}\theta_{td}}{\sum_s \phi_{ws}\theta_{sd}}.$$

M-шаг: частотные оценки условных вероятностей вычисляются путём суммирования счётчика $n_{dwt} = n_{dw}p(t|d,w)$:

$$\phi_{wt} = \frac{n_{wt}}{n_t}, \qquad n_{wt} = \sum_{d \in D} n_{dwt}, \qquad n_t = \sum_{w \in W} n_{wt};$$

$$\theta_{td} = \frac{n_{td}}{n_d}, \qquad n_{td} = \sum_{w \in d} n_{dwt}, \qquad n_d = \sum_{t \in T} n_{td}.$$

ARTM — аддитивная регуляризация тематической модели

Пусть, наряду с правдоподобием, требуется максимизировать ещё n критериев — *регуляризаторов* $R_i(\Phi, \Theta)$, $i = 1, \ldots, n$.

Метод многокритериальной оптимизации — *скаляризация*.

Задача максимизации регуляризованного правдоподобия:

$$\underbrace{\sum_{d \in D} \sum_{w \in d} n_{dw} \ln \sum_{t \in T} \phi_{wt} \theta_{td}}_{\mathscr{L}(\Phi,\Theta)} + \underbrace{\sum_{i=1}^{n} \tau_i R_i(\Phi, \Theta)}_{R(\Phi,\Theta)} \to \max_{\Phi,\Theta},$$

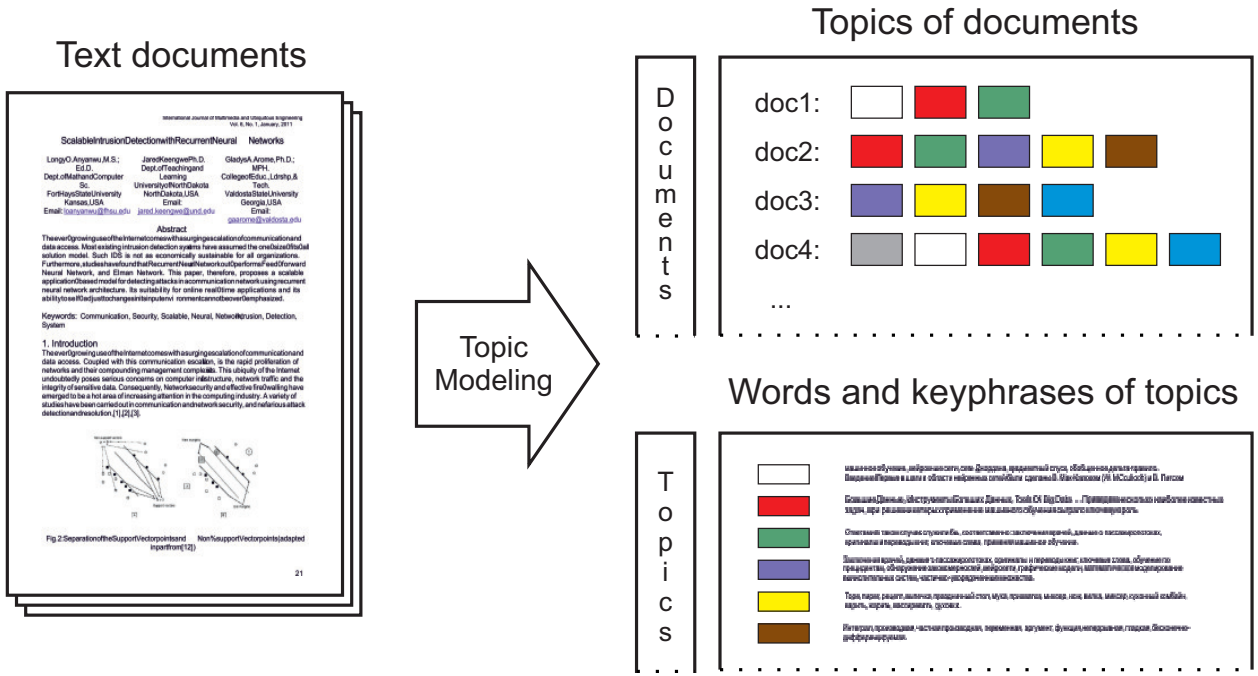
при ограничениях неотрицательности и нормировки

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

где $\tau_i > 0$ — *коэффициенты регуляризации*.

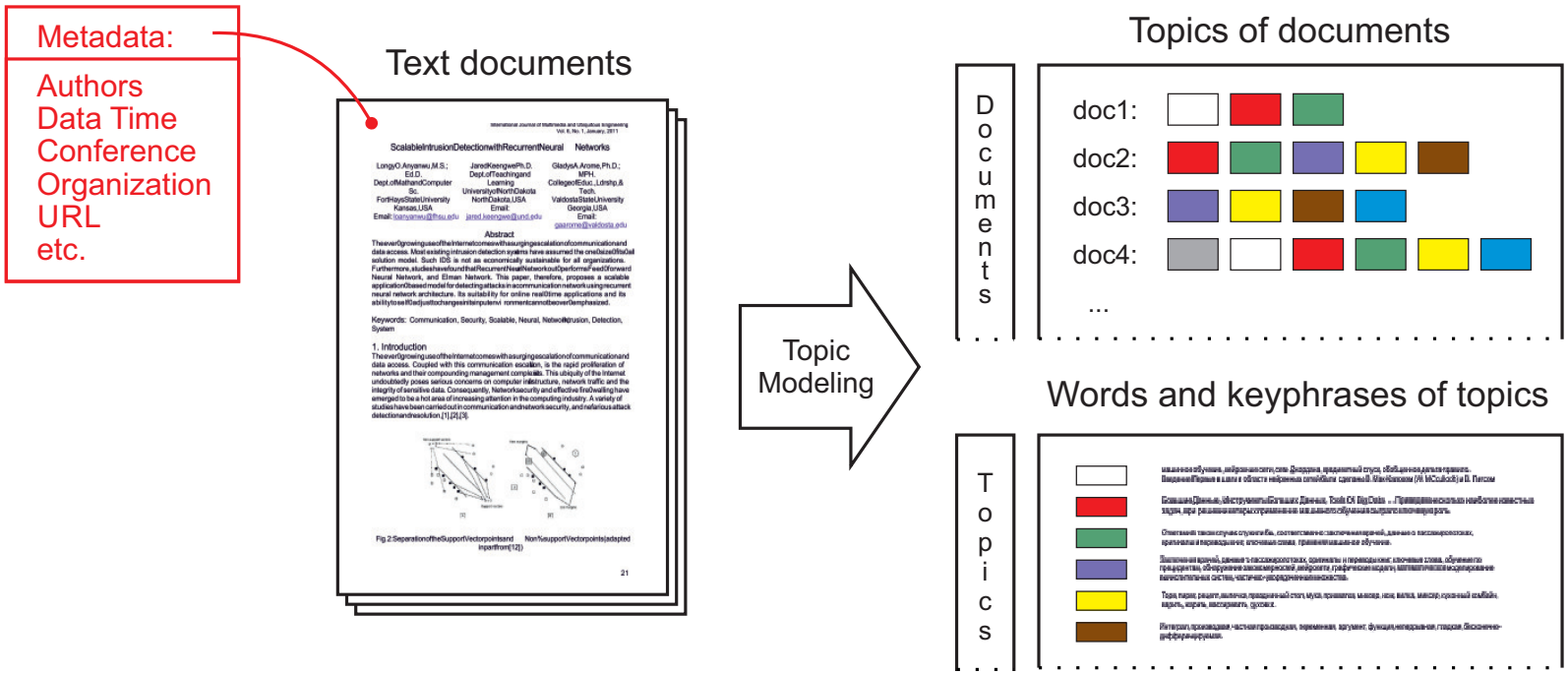
Мультимодальная тематическая модель

находит тематику документов $p(t|d)$, терминов $p(t|w)$,...



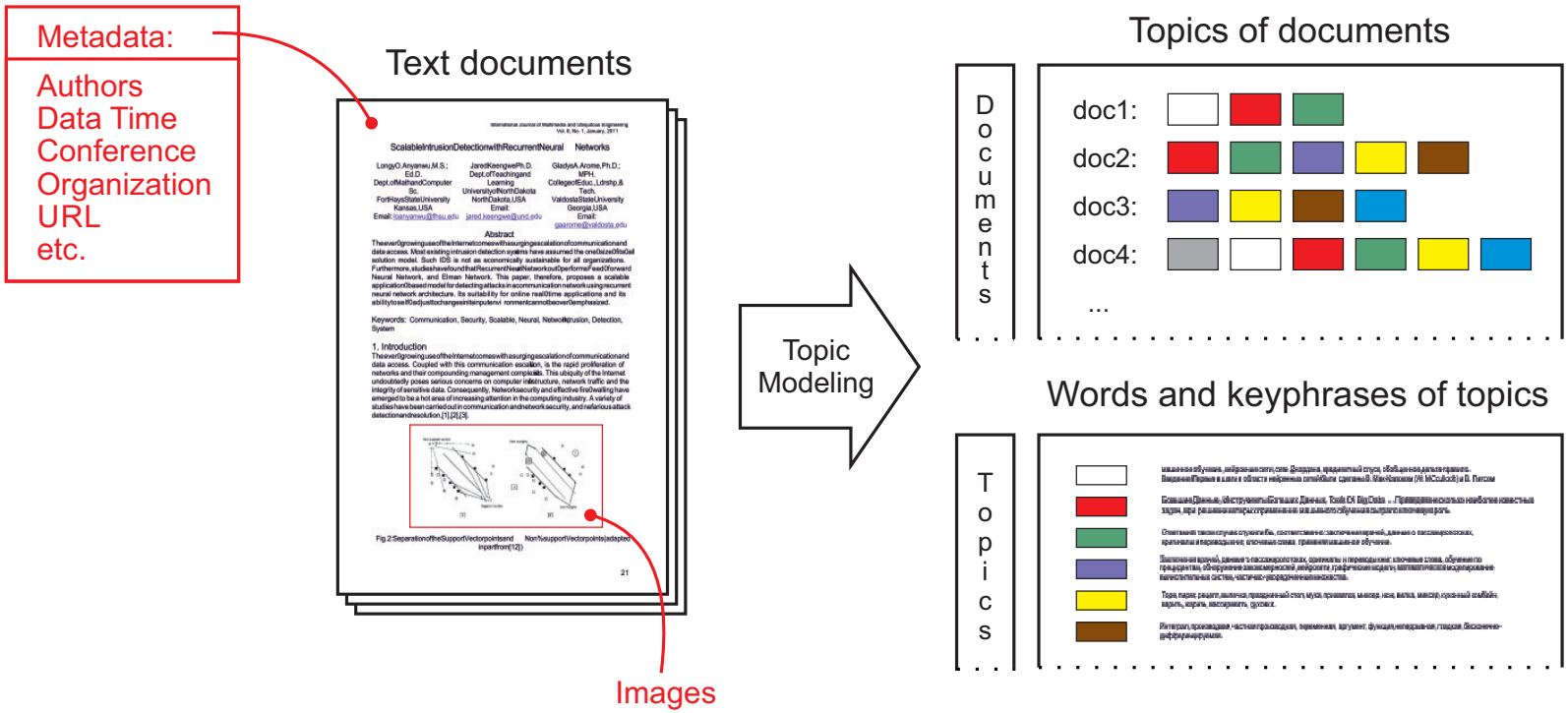
Мультимодальная тематическая модель

находит тематику документов $p(t|d)$, терминов $p(t|w)$,
авторов $p(t|a)$, времени $p(t|a)$,...



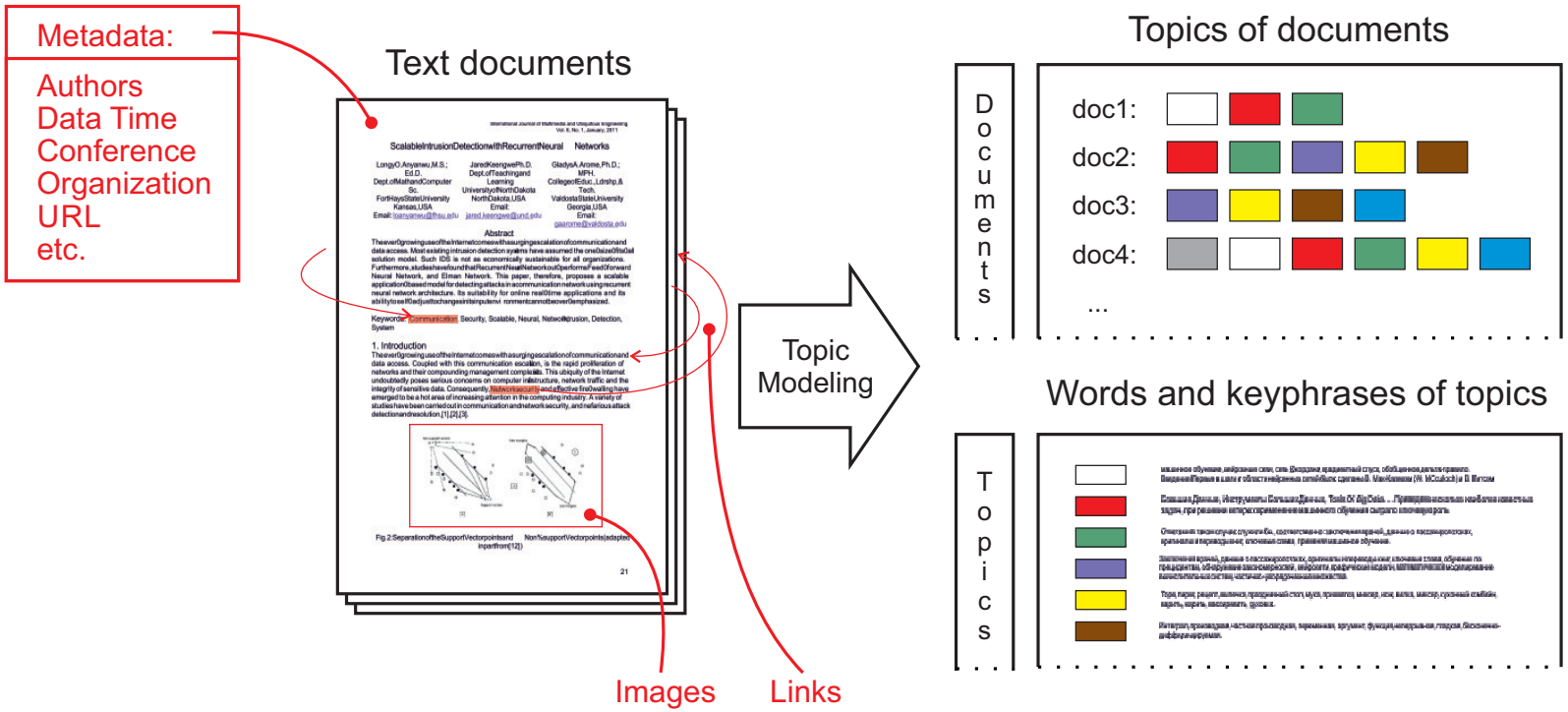
Мультимодальная тематическая модель

находит тематику документов $p(t|d)$, терминов $p(t|w)$,
авторов $p(t|a)$, времени $p(t|a)$, элементов изображений $p(t|e)$,...



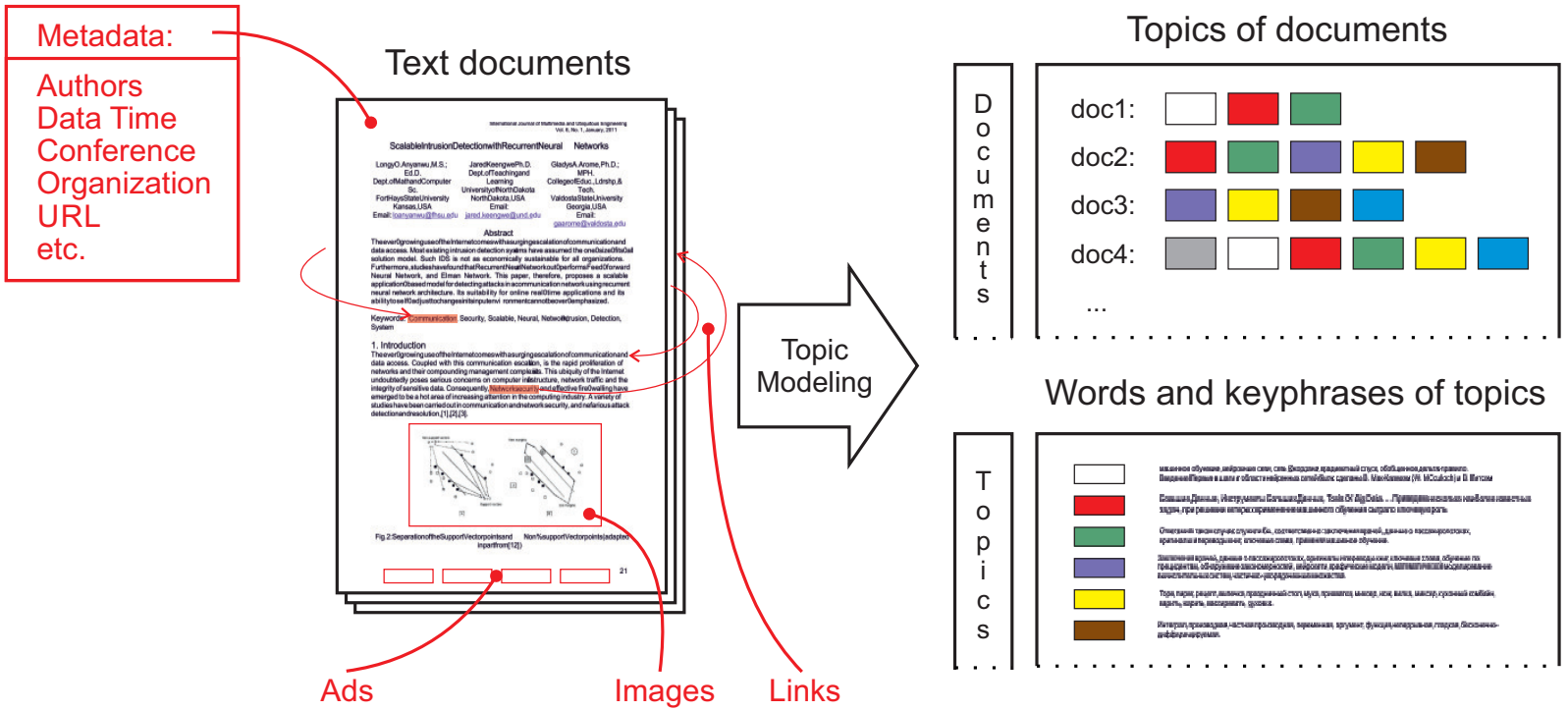
Мультимодальная тематическая модель

находит тематику документов $p(t|d)$, терминов $p(t|w)$, авторов $p(t|a)$, времени $p(t|a)$, элементов изображений $p(t|e)$, ссылок $p(d'|r)$,...



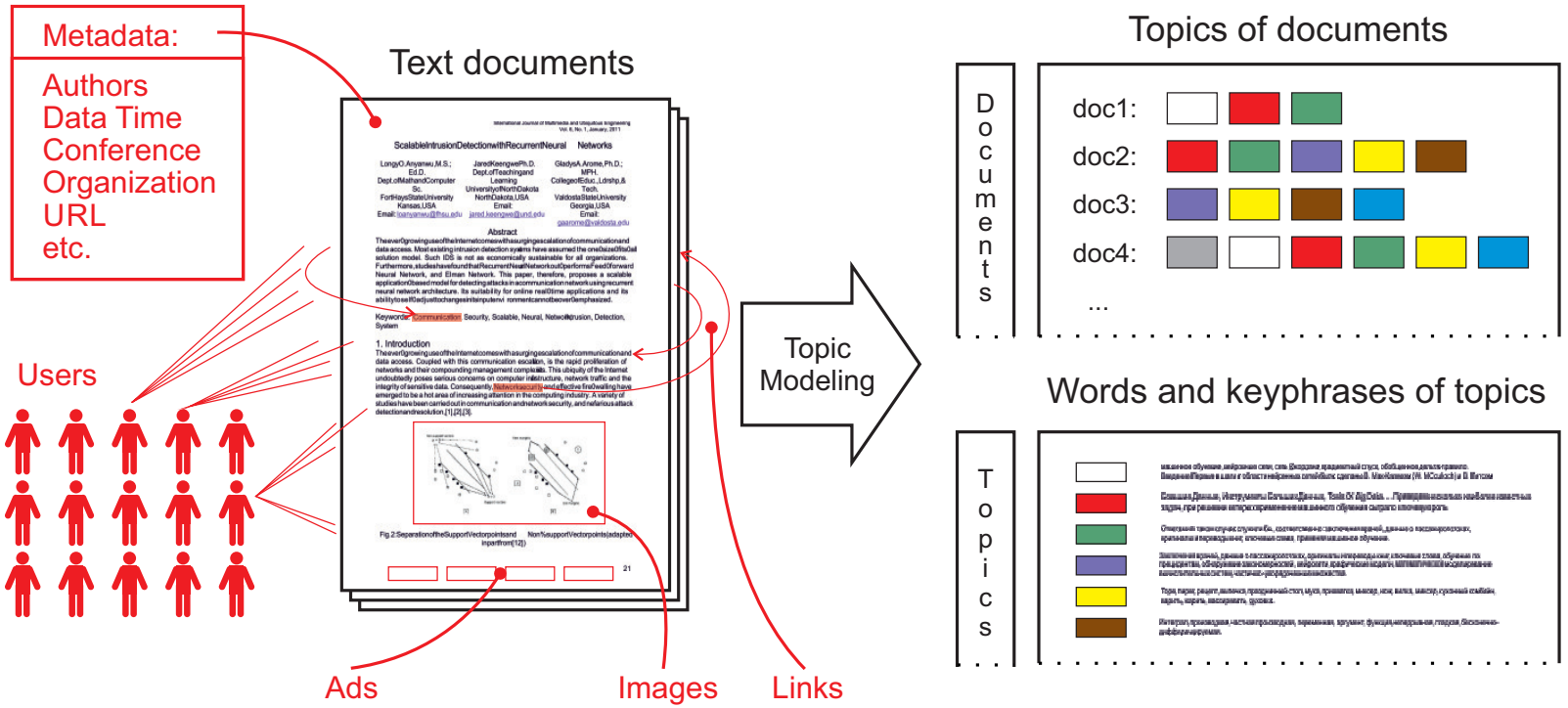
Мультимодальная тематическая модель

находит тематику документов $p(t|d)$, терминов $p(t|w)$, авторов $p(t|a)$, времени $p(t|a)$, элементов изображений $p(t|e)$, ссылок $p(d'|r)$, баннеров $p(t|b)$,...



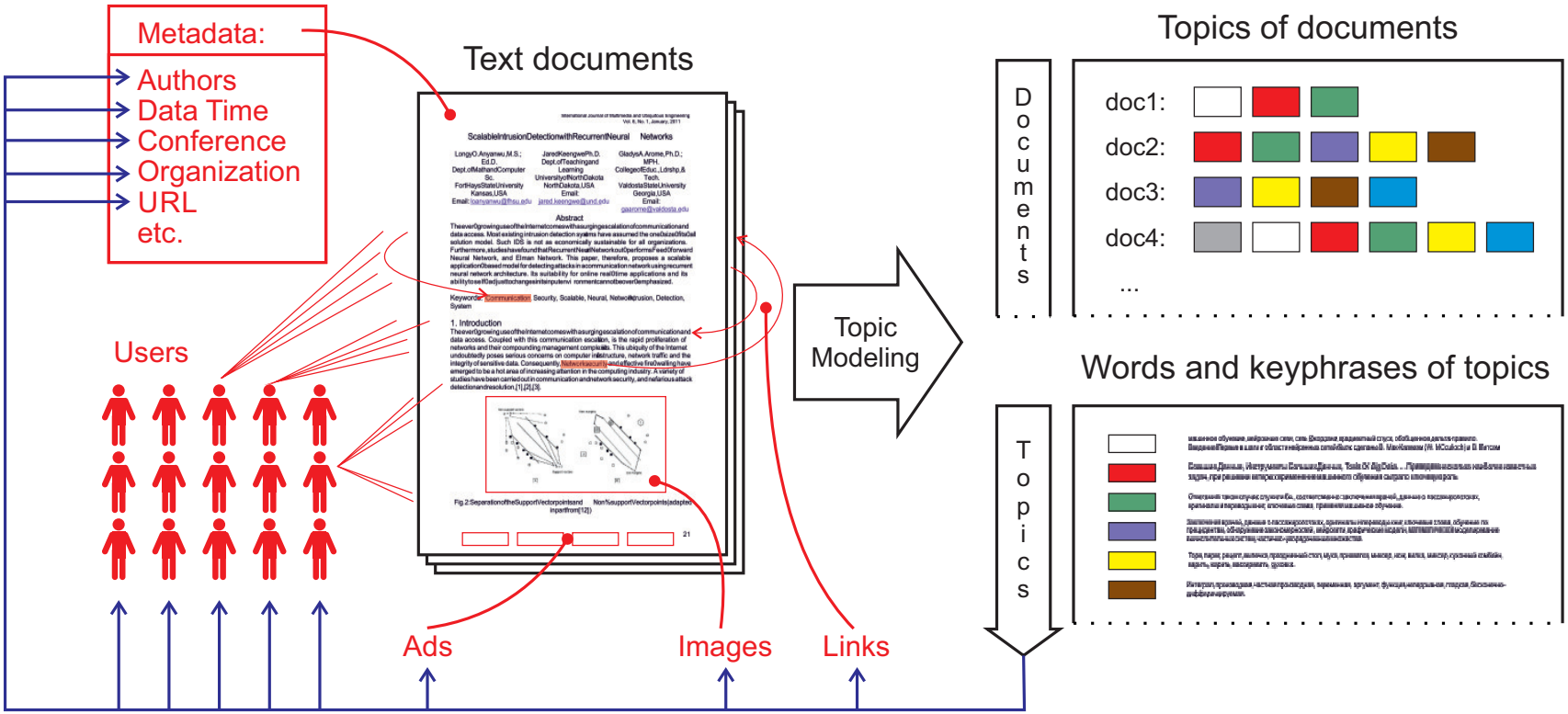
Мультимодальная тематическая модель

находит тематику документов $p(t|d)$, терминов $p(t|w)$, авторов $p(t|a)$, времени $p(t|a)$, элементов изображений $p(t|e)$, ссылок $p(d'|r)$, баннеров $p(t|b)$, пользователей $p(t|u)$,...



Мультимодальная тематическая модель

Каждая модальность $m \in M$ описывается своим словарём W^m, документы могут содержать элементы разных модальностей, каждая тема имеет своё распределение $p(w|t)$, $w \in W^m$



MultiARTM — мультимодальная ARTM

Каждая модальность $m \in M$ описывается своим словарём W^m, документы могут содержать элементы разных модальностей, каждая тема имеет своё распределение $p(w|t)$, $w \in W^m$

Задача максимизации регуляризованного правдоподобия:

$$\sum_{m\in M} \tau_m \underbrace{\sum_{d\in D}\sum_{w\in W^m} n_{dw} \ln \sum_{t\in T} \phi_{wt}\theta_{td}}_{\text{log-правдоподобие } \mathscr{L}_m(\Phi,\Theta)} + \underbrace{\sum_{i=1}^{n} \tau_i R_i(\Phi,\Theta)}_{R(\Phi,\Theta)} \to \max_{\Phi,\Theta},$$

при ограничениях неотрицательности и нормировки

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

где $\tau_m > 0$, $\tau_i > 0$ — *коэффициенты регуляризации*.

EM-алгоритм для мультимодальной ARTM

EM-алгоритм = метод простых итераций для системы уравнений

Теорема

Точка максимума $\mathscr{L}(\Phi,\Theta)+R(\Phi,\Theta)$ удовлетворяет системе уравнений со вспомогательными переменными $p_{tdw}=p(t|d,w)$:

$$p_{tdw} = \underset{t\in T}{\text{norm}}\left(\phi_{wt}\theta_{td}\right);$$

$$\phi_{wt} = \underset{w\in W^m}{\text{norm}}\left(n_{wt}+\phi_{wt}\frac{\partial R}{\partial \phi_{wt}}\right); \quad n_{wt}=\sum_{d\in D}\tau_{m(w)}n_{dw}p_{tdw};$$

$$\theta_{td} = \underset{t\in T}{\text{norm}}\left(n_{td}+\theta_{td}\frac{\partial R}{\partial \theta_{td}}\right); \quad n_{td}=\sum_{w\in d}\tau_{m(w)}n_{dw}p_{tdw};$$

где $m(w)$ — модальность термина w, т.е. $w\in W^{m(w)}$.

Доказательство Теоремы о регуляризации M-шага

1. Условия ККТ для ϕ_{wt}, $w \in W^m$ (для θ_{td} всё аналогично):

$$\sum_d \tau_m n_{dw} \frac{\theta_{td}}{p(w|d)} + \frac{\partial R}{\partial \phi_{wt}} = \lambda_t - \lambda_{wt}; \quad \lambda_{wt} \geqslant 0; \quad \lambda_{wt}\phi_{wt} = 0.$$

2. Умножим обе части равенства на ϕ_{wt} и выделим p_{tdw}:

$$\phi_{wt}\lambda_t = \sum_d \tau_m n_{dw} \frac{\phi_{wt}\theta_{td}}{p(w|d)} + \phi_{wt}\frac{\partial R}{\partial \phi_{wt}} = n_{wt} + \phi_{wt}\frac{\partial R}{\partial \phi_{wt}}.$$

3. Альтернатива: либо $\phi_{wt} = 0$ для всех w, либо $\lambda_t > 0$ и

$$\phi_{wt}\lambda_t = \left(n_{wt} + \phi_{wt}\frac{\partial R}{\partial \phi_{wt}}\right)_+.$$

4. Суммируем обе части равенства по $w \in W^m$:

$$\lambda_t = \sum_{w \in W^m} \left(n_{wt} + \phi_{wt}\frac{\partial R}{\partial \phi_{wt}}\right)_+.$$

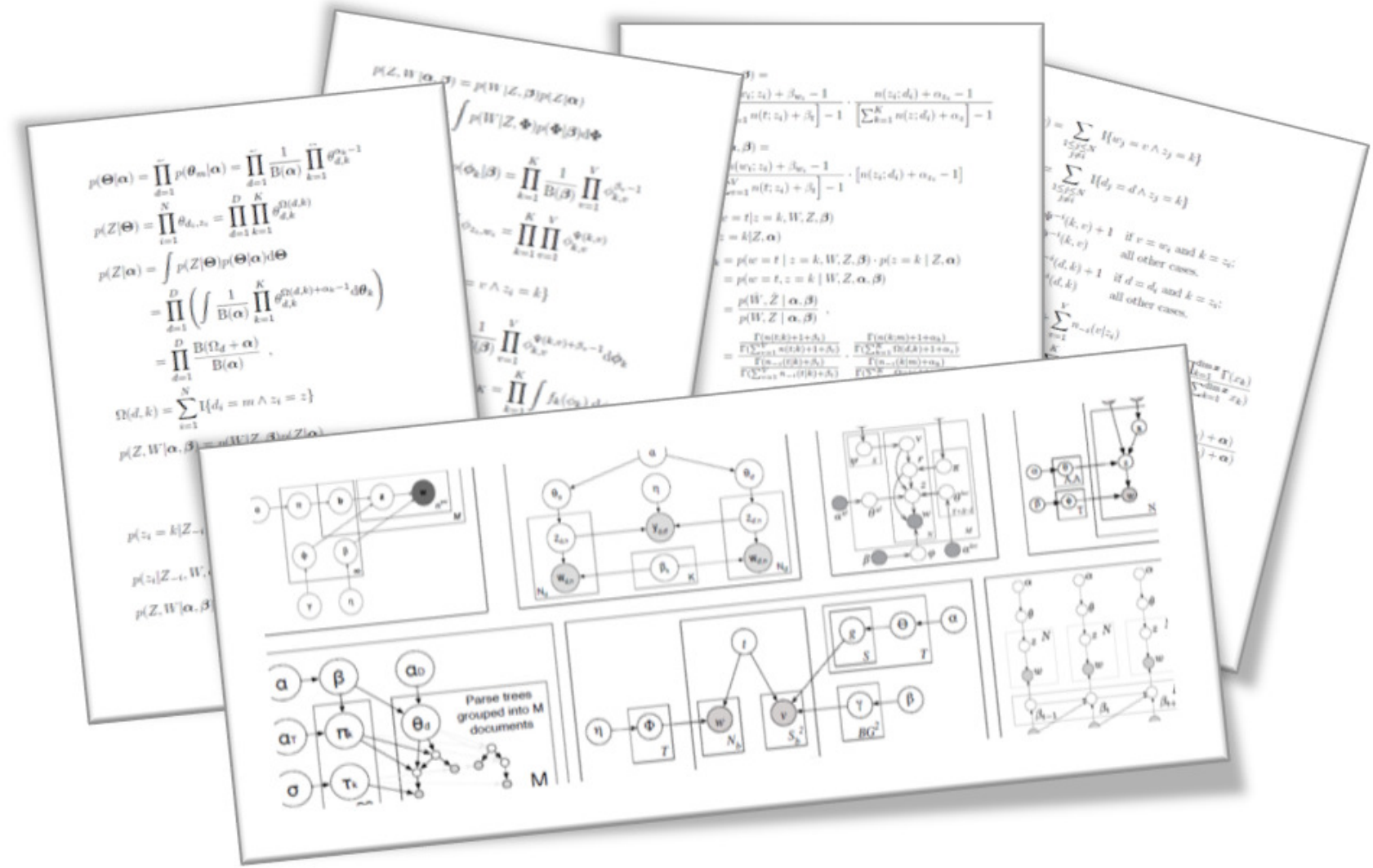
5. Подставим λ_t из (4) в (3), получим требуемое. ■

ARTM: зоопарк регуляризаторов

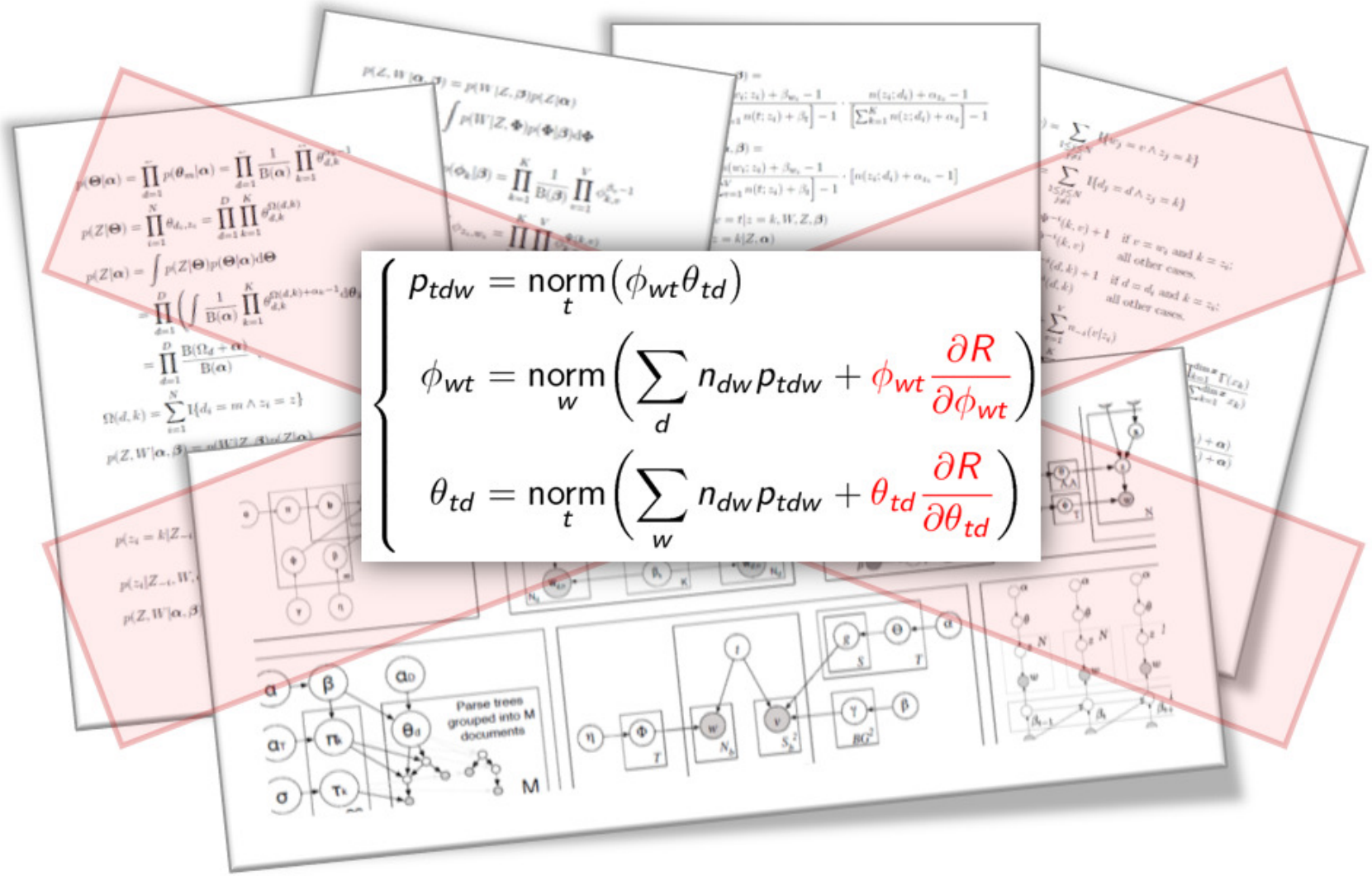
- сглаживание тем общей лексики (LDA)
- разреживание предметных тем
- декоррелирование предметных тем
- энтропийное разреживание для отбора тем
- максимизация согласованности (когерентности)
- обучение с учителем для классификации и регрессии
- частичное (semi-supervised) обучение
- динамическое (темпоральное) моделирование
- многоязычное тематическое моделирование
- и др.

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

ARTM — альтернатива байесовскому обучению



ARTM — альтернатива байесовскому обучению



BigARTM: библиотека тематического моделирования

Ключевые возможности:

- Онлайновый параллельный MultiARTM
- Большие данные: коллекция не хранится в памяти
- Встроенная библиотека регуляризаторов

BigARTM: библиотека тематического моделирования

Ключевые возможности:

- Онлайновый параллельный MultiARTM
- Большие данные: коллекция не хранится в памяти
- Встроенная библиотека регуляризаторов

Сообщество:

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

BigARTM: библиотека тематического моделирования

Ключевые возможности:

- Онлайновый параллельный MultiARTM
- Большие данные: коллекция не хранится в памяти
- Встроенная библиотека регуляризаторов

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

Эксперимент 1. Обгоняем конкурентов по скорости

- 3.7M статей английской Вики, 100K уникальных слов

	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* = число параллельных потоков
- *inference* = время тематизации 100K тестовых документов
- *perplexity* вычислена на тестовой выборке документов

Эксперимент 2. Мультиязычная модель

Модальности — это разные языки.

216 175 русско-английских пар статей Вики.
Первые 10 слов и их вероятностями $p(w|t)$ в %:

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

Эксперимент 2. Мультиязычная модель

216 175 русско-английских пар статей Вики.
Первые 10 слов и их вероятностями $p(w|t)$ в %:

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

Независимый асессор оценил 396 тем из $|T| = 400$ как хорошо интерпретируемые.

Эксперимент 3. Интерпретируемость мультиграммной модели

Две модальности — униграммы и биграммы.

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

Энтропийный разреживающий регуляризатор для отбора тем

Чтобы сделать распределение $p(t)$ разреженным, максимизируем его KL-дивергенцию с равномерным распределением: $\mathrm{KL}\left(\frac{1}{|T|} \,\|\, p(t)\right) \to \max$:

$$R(\Theta) = -\tau n \sum_{t \in S} \frac{1}{|T|} \ln \underbrace{\sum_{d \in D} p(d)\theta_{td}}_{p(t)} \to \max.$$

Регуляризованный M-шаг разреживает строки Θ целиком:

$$\theta_{td} = \underset{t \in T}{\mathrm{norm}} \left(n_{td} \left(1 - \tau \frac{n}{n_t |T|} \right) \right).$$

Если $n_t < \tau \frac{n}{|T|}$, то все элементы t-й строки обращаются в нуль.

Эксперименты с энтропийным разреживанием тем

Коллекция статей NIPS (Neural Information Processing System)

- $|D| = 1566$ обучающих документов; $|D'| = 174$ тестовых
- $|W| \approx 1.3 \cdot 10^4$ — мощность словаря

Синтетическая коллекция:

- строим PLSA за 500 итераций, $|T_0| = 50$ тем на NIPS
- генерируем (n^0_{dw}) из полученных Φ и Θ:

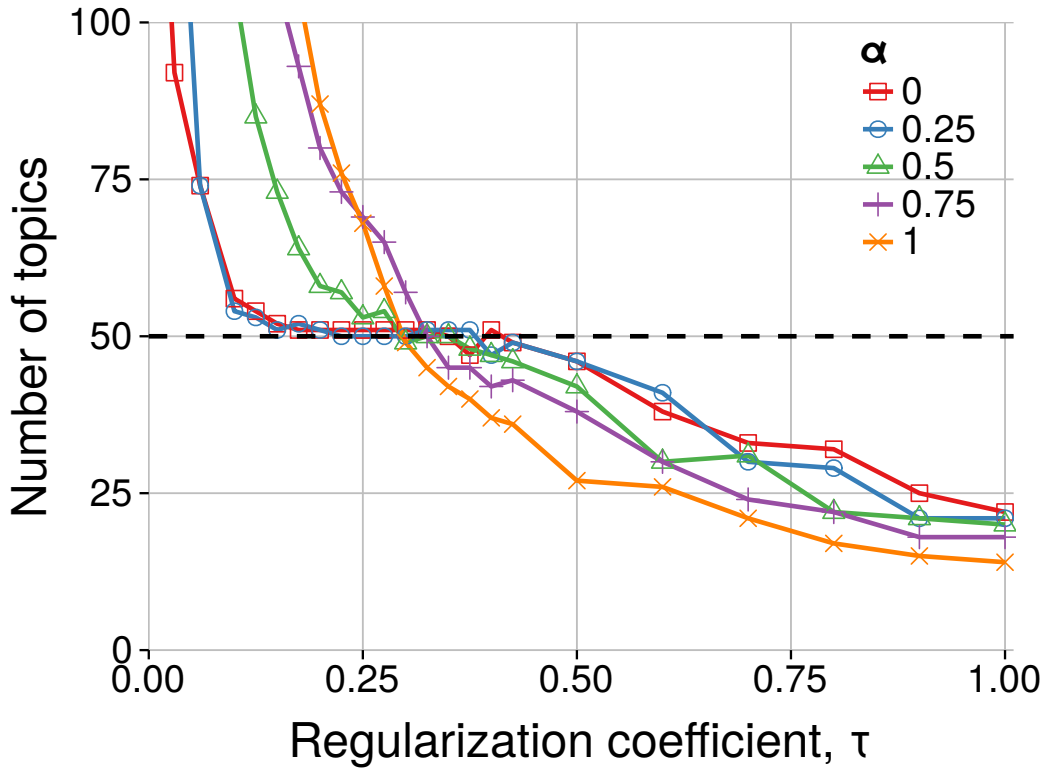
$$n^0_{dw} = n_d \sum_{t \in T} \phi_{wt} \theta_{td}$$

Параметрическое семейство полусинтетических данных:

- n^α_{dw} — смесь синтетических данных n^0_{dw} и реальных n_{dw}:

$$n^\alpha_{dw} = \alpha n_{dw} + (1 - \alpha) n^0_{dw}$$

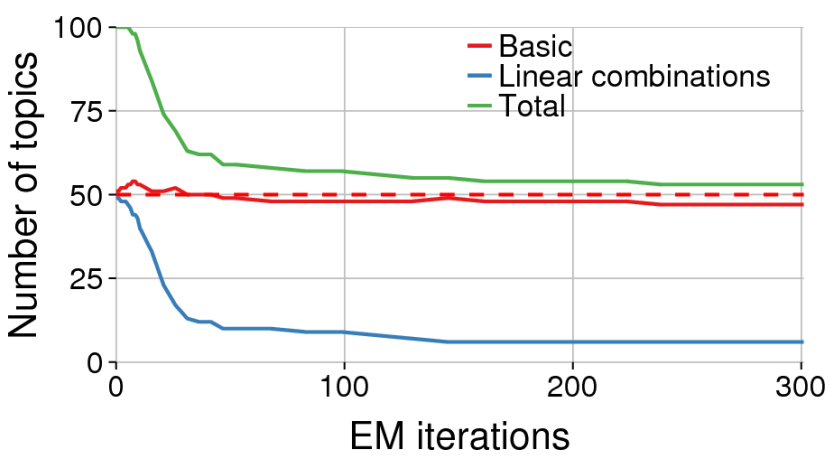
Попытка определения числа тем

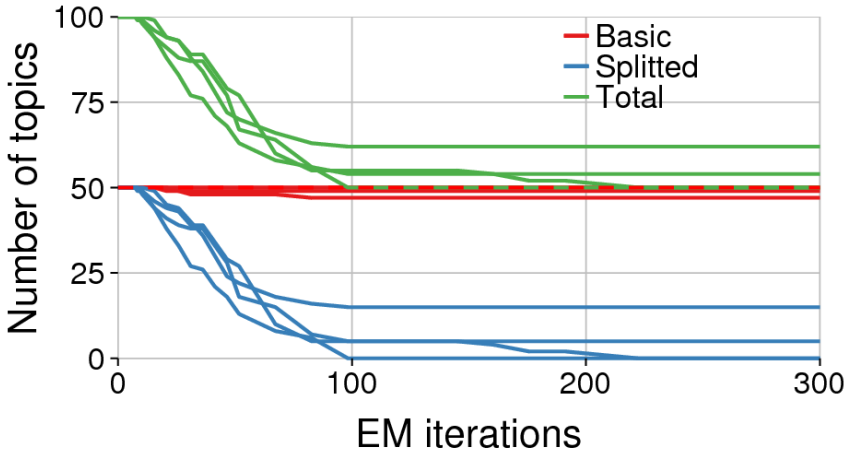


- На синтетических данных надёжно находим $|T| = 50$,
- в широком интервале значений коэффициента τ;
- однако на реальных данных нет столь чёткого интервала.

Удаление линейно зависимых и расщеплённых тем

- Добавили 50 линейных комбинаций тем в модельную Φ.
- Расщепили 50 тем, каждую на две подтемы в модельной Φ.





- Удаляются линейно зависимые и расщеплённые темы
- Остаются более различные темы исходной модели.

Регуляризаторы для динамической тематической модели

Y — моменты времени (например, годы публикаций),
$y(d)$ — метка времени документа d,
$D_y \subset D$ — все документы, относящиеся к моменту $y \in Y$.

Гипотеза 1: распределение $p(t|y) = \sum_{d \in D_y} \theta_{td} p(d)$ разрежено:

$$R_1(\Theta) = -\tau_1 \sum_{y \in Y} \mathrm{KL}\left(\tfrac{1}{|T|} \| p(t|y)\right) \to \max.$$

Гипотеза 2: $p(y|t)$ меняются плавно, с редкими скачками:

$$R_2(\Theta) = -\tau_2 \sum_{y \in Y} \sum_{t \in T} \left|p(y|t) - p(y{-}1|t)\right| \to \max.$$

Эксперименты. Задача анализа потока пресс-релизов

Коллекция официальных пресс-релизов внешнеполитических ведомств ряда стран на английском языке.
Более 20 тыс. сообщений за 10 лет, 180Мб текста.

Найти:

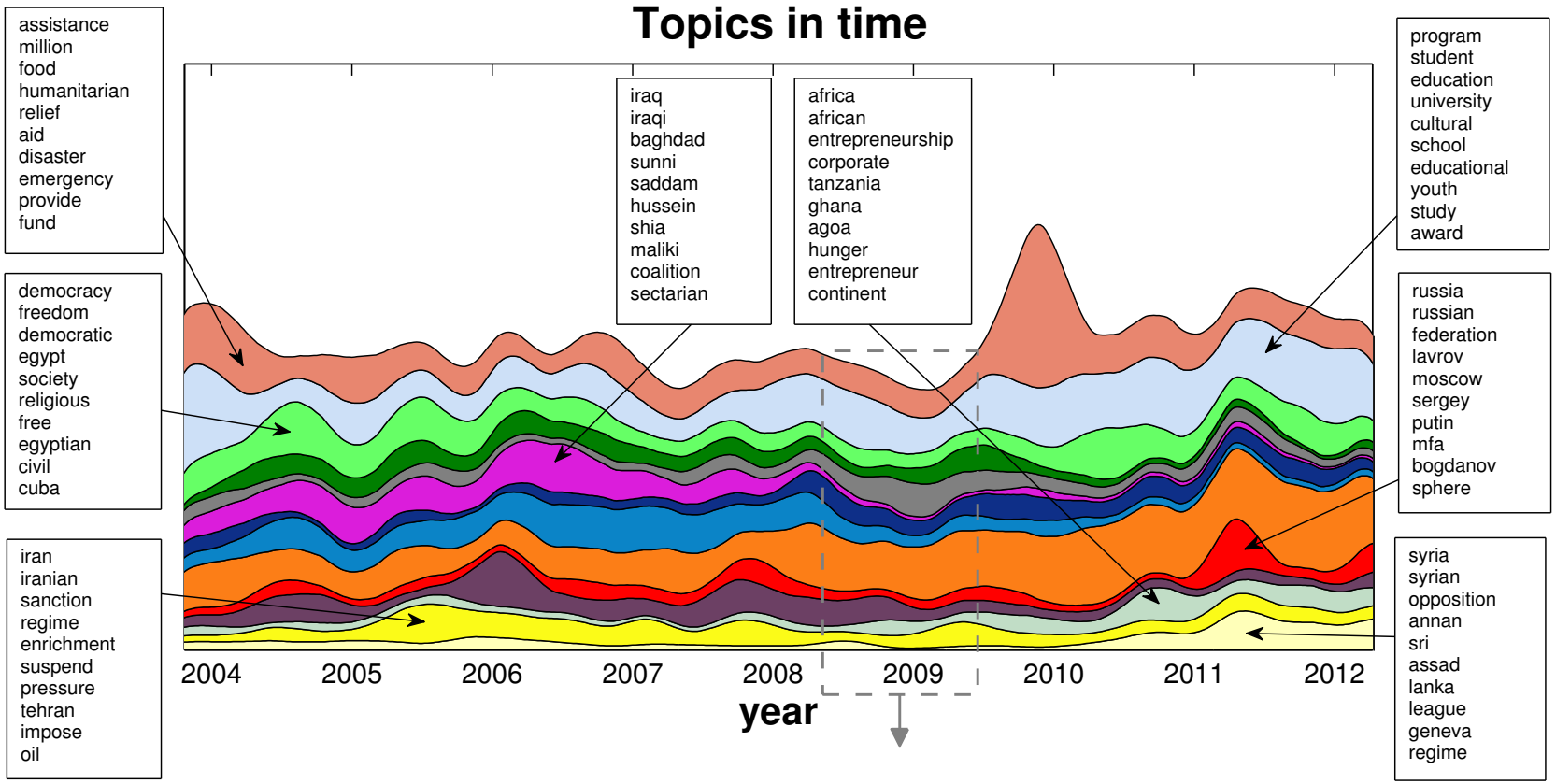
- какие темы перманентные?
- какие темы привязаны к событиям?
- какие темы и в какие моменты коррелируют?

Регуляризаторы:

- разреживание, сглаживание, декоррелирование
- разреживание тем $p(t|y)$ в каждый момент времени y
- сглаживание тем $p(y|t)$ в соседние моменты времени

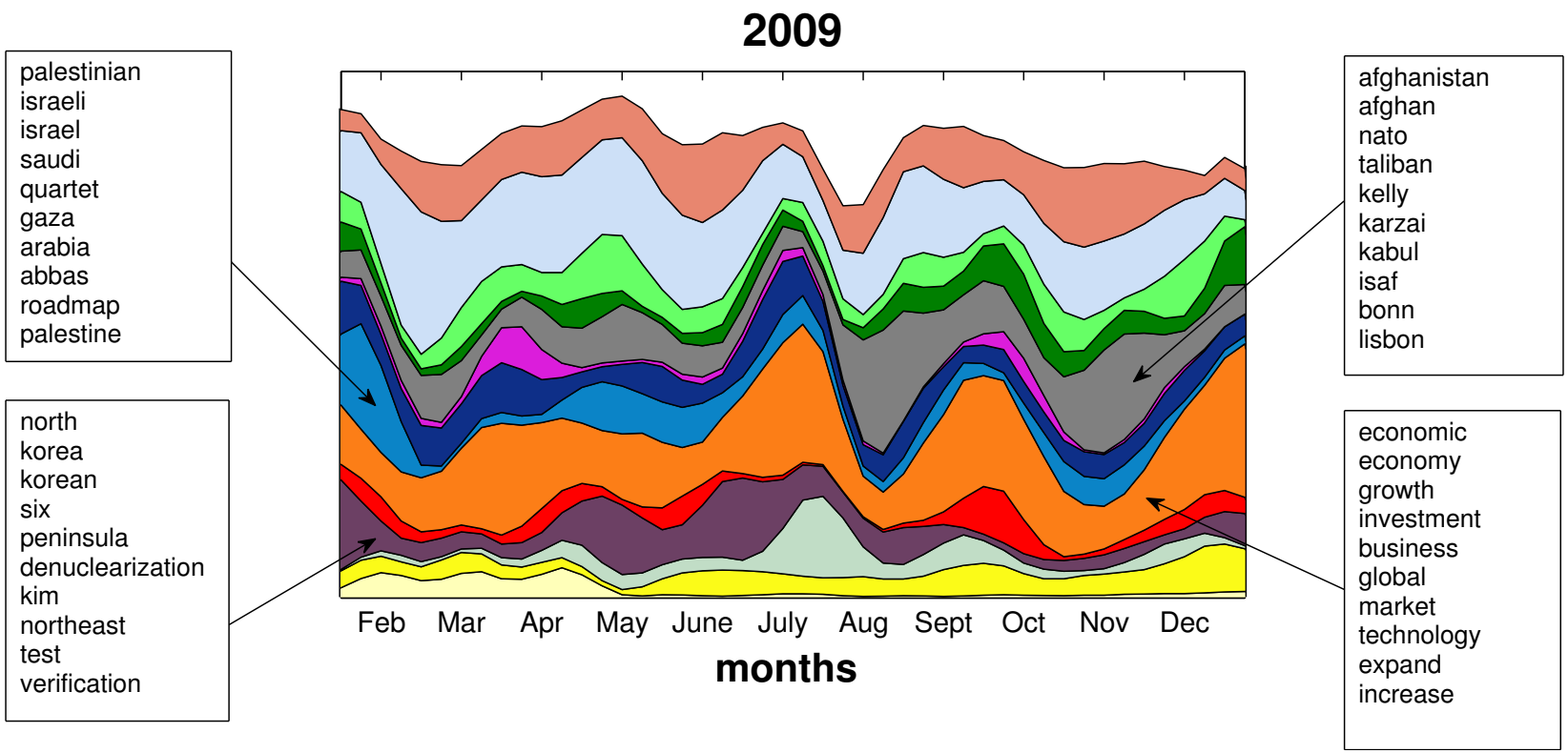
Эксперименты. Задача анализа потока пресс-релизов

Примеры хорошо интерпретируемых тем



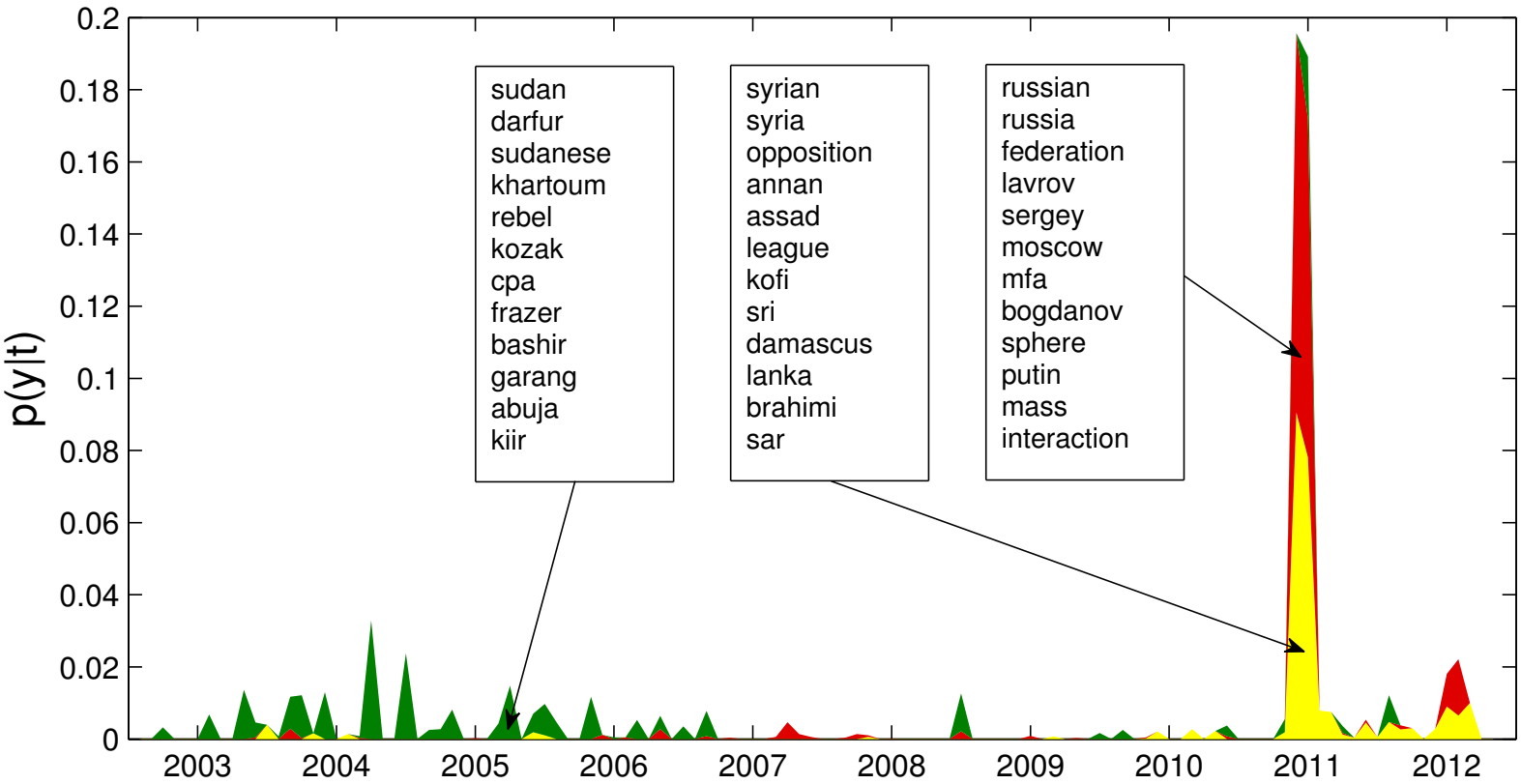
Эксперименты. Задача анализа потока пресс-релизов

Укрупнение масштаба времени



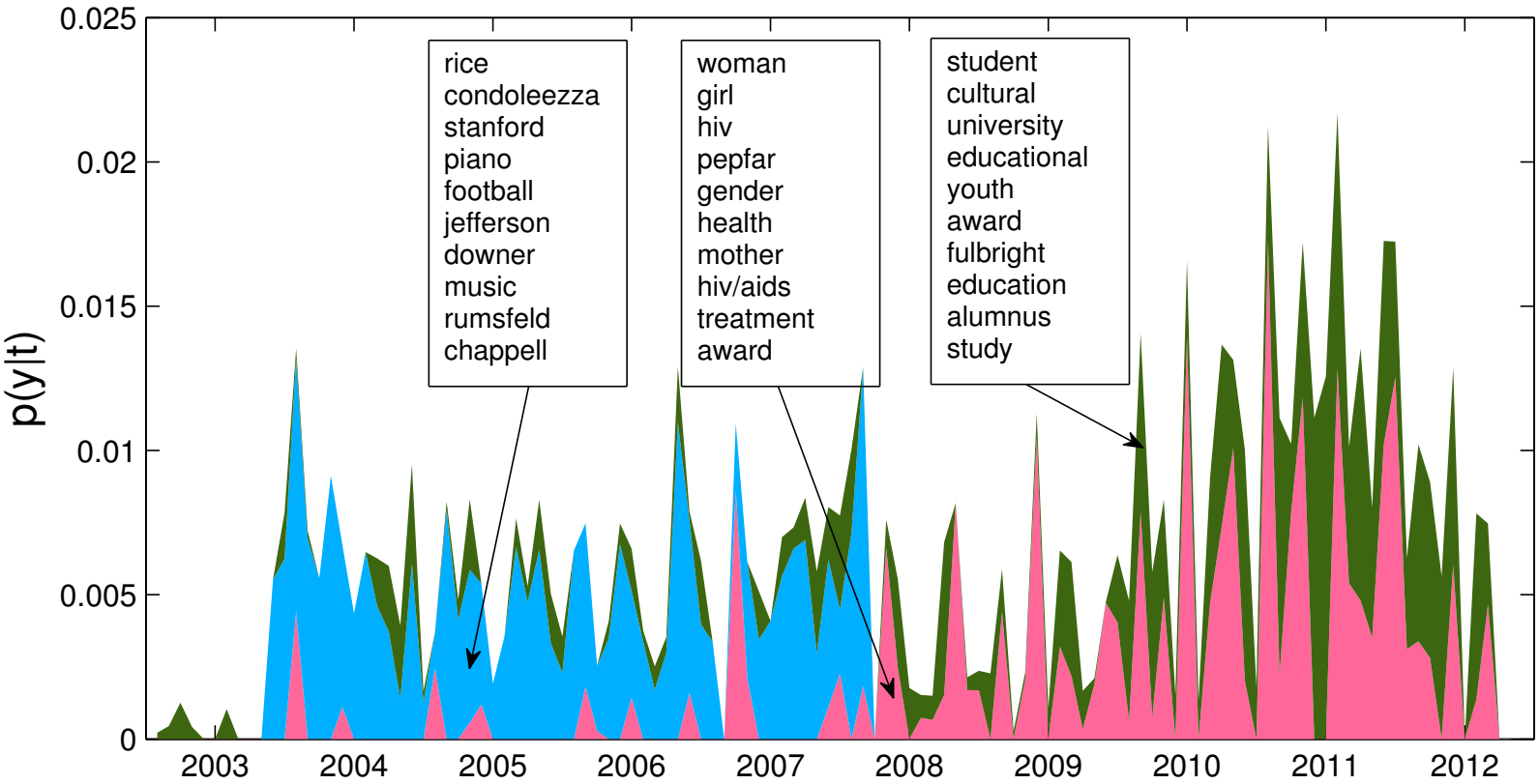
Эксперименты. Задача анализа потока пресс-релизов

Примеры событийных тем и момента их совместного всплеска



Эксперименты. Задача анализа потока пресс-релизов

Примеры перманентных тем



Мотивации

Выборка может содержать не только пары (d, w), но также тройки, ..., n-ки элементов разных модальностей.

Примеры:

- **Данные социальной сети:**
  (d, u, w) — в блоге d пользователь u написал слово w
- **Данные сети интернет-рекламы:**
  (u, d, b) — пользователь u кликнул рекламное объявление b на веб-странице d
- **Данные рекомендательной системы:**
  (u, f, s) — пользователь u оценил фильм f в ситуативном контексте s

Хотим объяснить наблюдаемую выборку рёбер гиперграфа латентными тематическими профилями его вершин.

Тематическая модель гиперграфа: определения и обозначения

$\Gamma = \langle V, E \rangle$ — ориентированный гиперграф

M — множество модальностей
$\mu\colon V \to M$, $\mu(v)$ — модальность вершины v
V^m — множество всех вершин модальности m
K — множество типов рёбер,
$\kappa\colon E \to K$, $\kappa(e)$ — тип ребра e

$h = h(k)$ — размерность рёбер типа k
$k \mapsto (m_0, \ldots, m_h) \in M^{h+1}$ — модальности вершин рёбер типа k
m_0 — выделенная модальность-контейнер ($\sim$ документ)
$e = (d, x)$, $x = (v_1, \ldots, v_h)$ — ребро с вершиной-контейнером d

X^k — наблюдаемая выборка рёбер типа k
n_{dx} — число вхождений ребра $e = (d, x)$ в выборку X^k
$p_k(d, x)$ — неизвестное распределение на рёбрах типа k

Тематическая модель гиперграфа

Вероятностная тематическая модель рёбер типа k:

$$p_k(x|d) = \sum_{t \in T} \theta_{td} \prod_{v \in x} \phi_{kvt},$$

$\theta_{td} = p(t|d)$ — не зависят от k
$\phi_{kvt} = p_k(v|t)$ — распределение элементов модальности $\mu(v)$ в теме t на рёбрах типа k

Задача максимизации регуляризованного правдоподобия:

$$\sum_{k \in K} \tau_k \sum_{(d,x) \in X^k} n_{dx} \ln \sum_{t \in T} \theta_{td} \prod_{v \in x} \phi_{kvt} + R(\Phi, \Theta) \to \max_{\Phi, \Theta},$$

$$\phi_{kvt} \geqslant 0, \quad \sum_{v \in V^m} \phi_{kvt} = 1, \; k \in K; \qquad \theta_{td} \geqslant 0, \quad \sum_{t \in T} \theta_{td} = 1,$$

где $\tau_k > 0$ — веса типов рёбер.

EM-алгоритм для мультимодальной ARTM

EM-алгоритм = метод простых итераций для системы уравнений

Теорема

Точка максимума удовлетворяет системе уравнений со вспомогательными переменными $p_{tdx} = p(t|d,x)$:

$$p_{tdx} = \underset{t\in T}{\text{norm}}\left(\theta_{td}\prod_{v\in x}\phi_{kvt}\right);$$

$$\phi_{kvt} = \underset{v\in V^m}{\text{norm}}\left(n_{kvt} + \phi_{kvt}\frac{\partial R}{\partial \phi_{kvt}}\right); \quad n_{kvt} = \sum_{(d,x)\in X^k}\big[v\in x\big]\tau_k n_x p_{tdx};$$

$$\theta_{td} = \underset{t\in T}{\text{norm}}\left(n_{td} + \theta_{td}\frac{\partial R}{\partial \theta_{td}}\right); \quad n_{td} = \sum_{(d,x)\in X^k}\tau_k n_x p_{tdx};$$

Направления дальнейших исследований (тосты)

- За интерпретируемость тем!
- За большое число тем!
- За мелкозернистую иерархию!
- За гиперграфовые обобщения!
- За лингвистическую регуляризацию!
- За автоматическое именование тем!
- За развитие BigARTM!

http://bigartm.org

Join BigARTM community!