Тематическое моделирование структуры расходов клиентов банка

Воронцов Константин Вячеславович
(ФИЦ ИУ РАН • МФТИ • МГУ • ВШЭ • Яндекс)

Айсина Роза Мунеровна
(ВМК МГУ)

Москва • 12 ноября 2016

Аннотация

Вероятностное тематическое моделирование обычно применяется для информационного поиска, классификации и анализа документов в больших текстовых коллекциях.

В его основе лежат весьма универсальные математические методы — матричные разложения и регуляризация некорректно поставленных задач.

Поэтому спектр применений тематического моделирования может выходить далеко за рамки текстовой аналитики.
Мы применили его для выявления латентных паттернов в структуре потребления клиентов Сбербанка.

Содержание

Что такое «тема» в коллекции текстовых документов?

Неформально,

- *тема* — семантически однородный кластер текстов
- *тема* — специальная терминология предметной области
- *тема* — набор часто совместно встречающихся терминов
- тем много меньше, чем терминов и чем документов

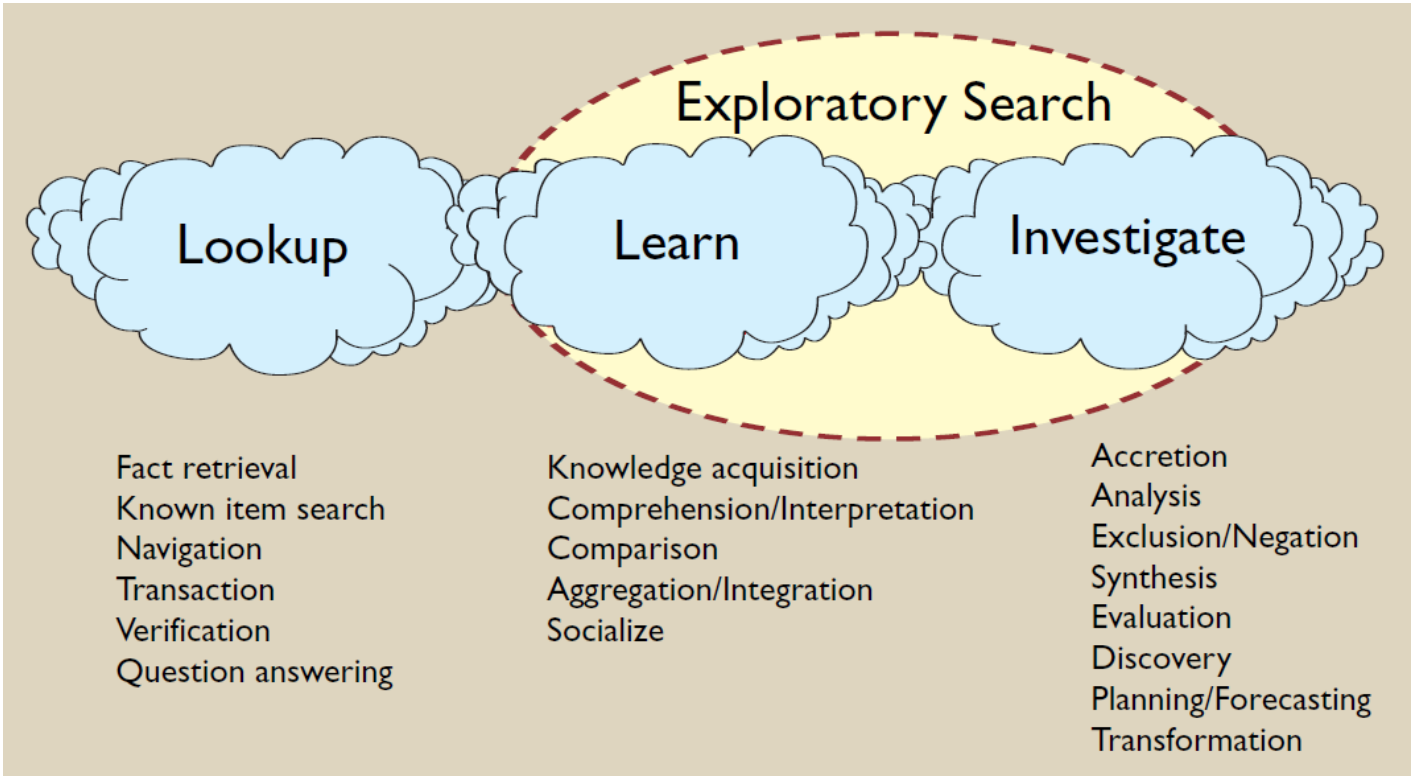
Более формально,

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

Тематическая модель оценивает вероятности $p(w|t)$ и $p(t|d)$ по наблюдаемым частотам $p(w|d)$ слов w в документах d.

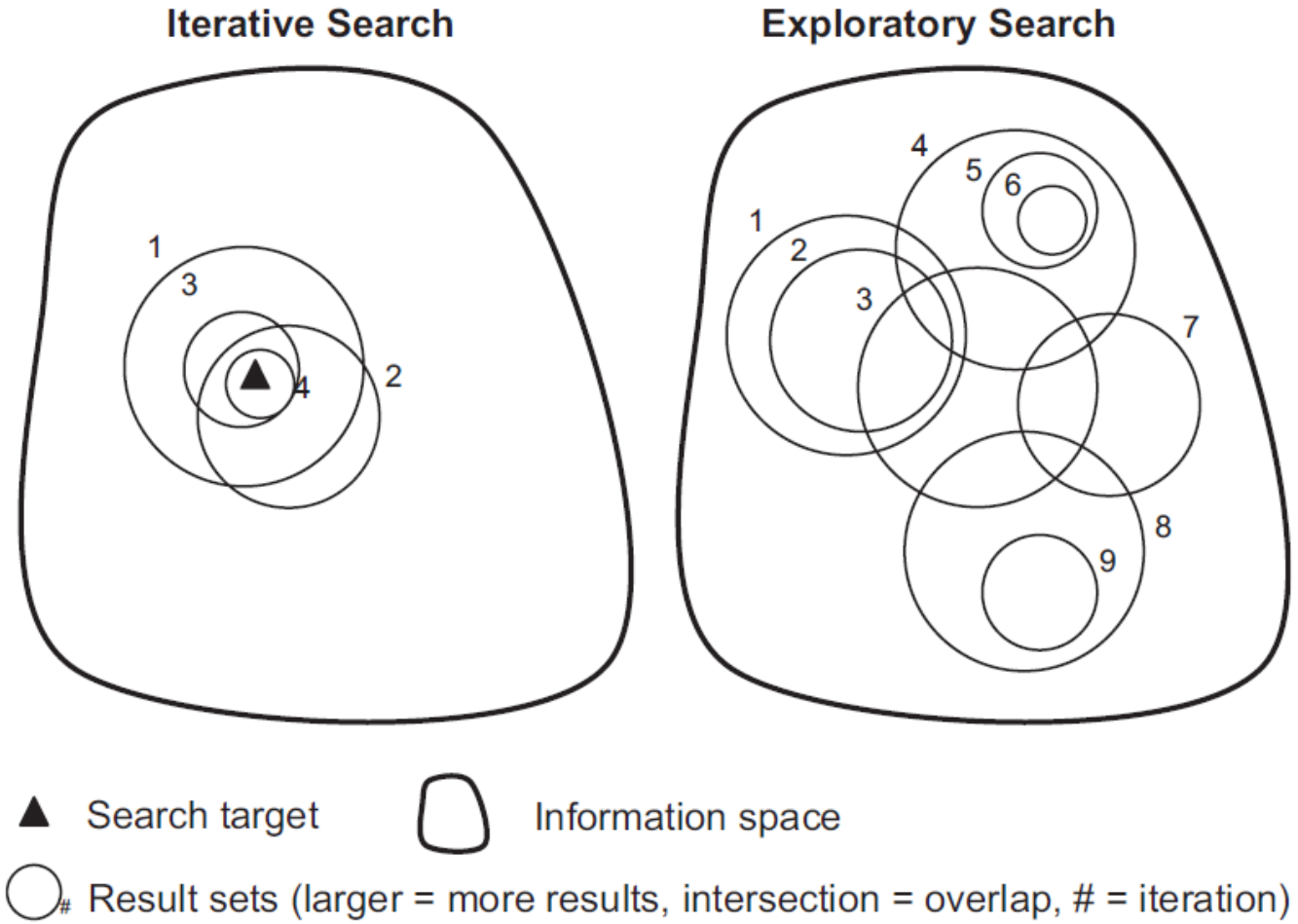
Концепция разведочного поиска (exploratory search)

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



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

От итераций «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

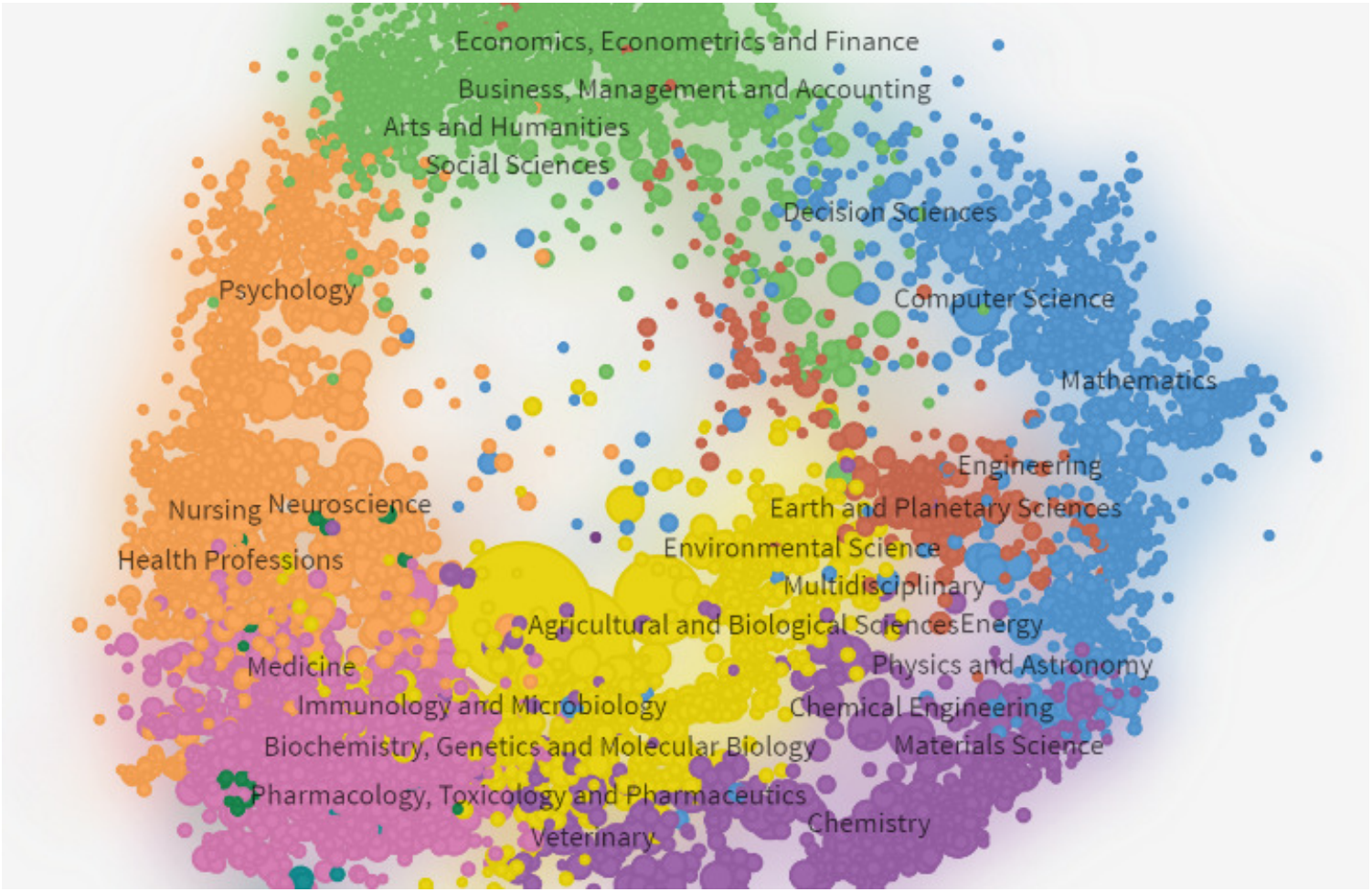
Цели поиска:

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

Сценарий поиска:

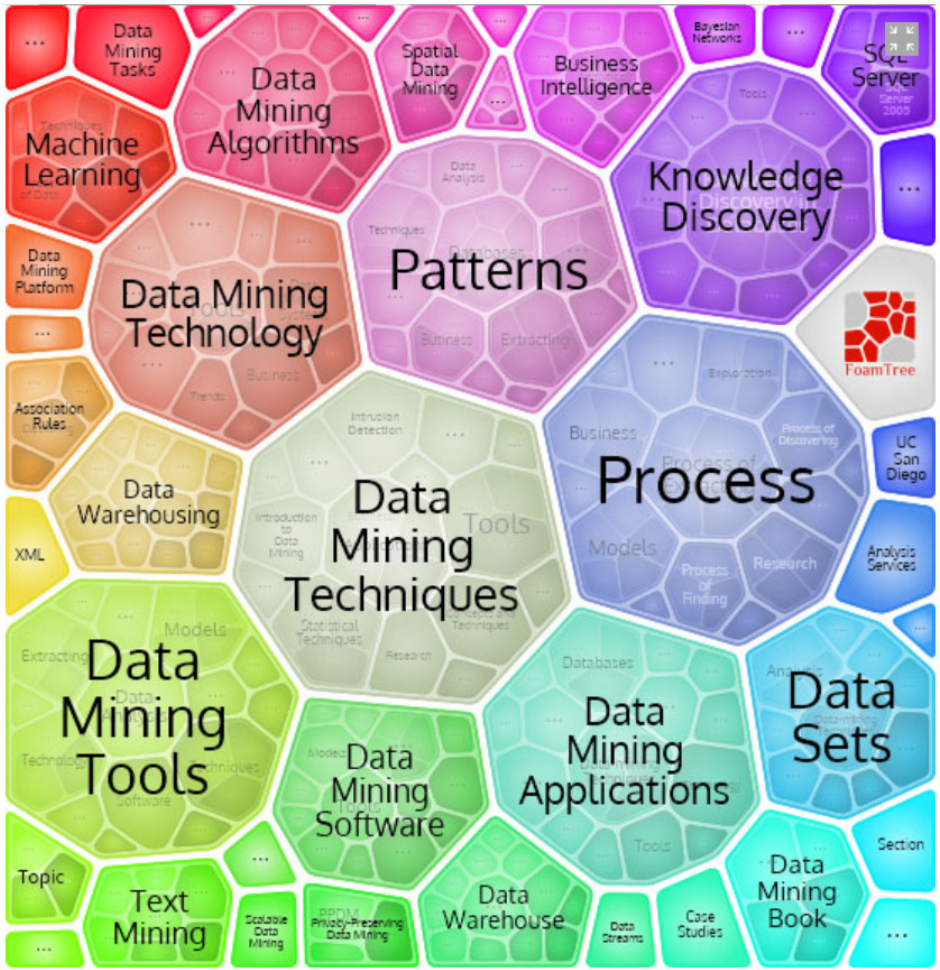
1. имея любой текст под рукой, в любом приложении,
2. хотим получить картину содержащихся в нём тем-подтем,
3. и «дорожную карту» предметной области в целом

Пример карты науки



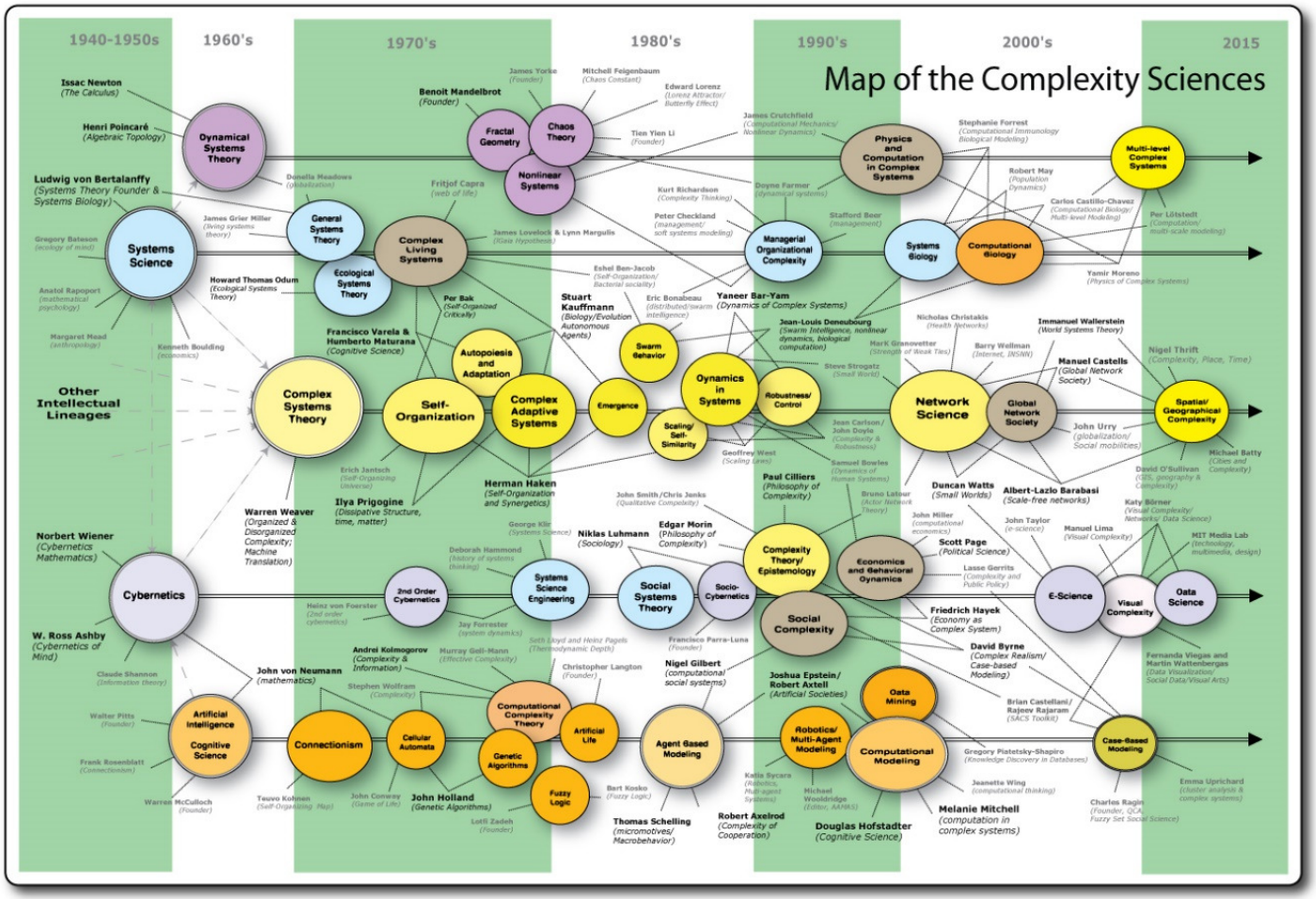
http://onlinelibrary.wiley.com/browse/subjects

Пример иерархической карты области *Data Mining*



FoamTree: https://carrotsearch.com/foamtree

Пример карты предметной области, построенной вручную



http://www.theoryculturesociety.org/brian-castellani-on-the-complexity-sciences

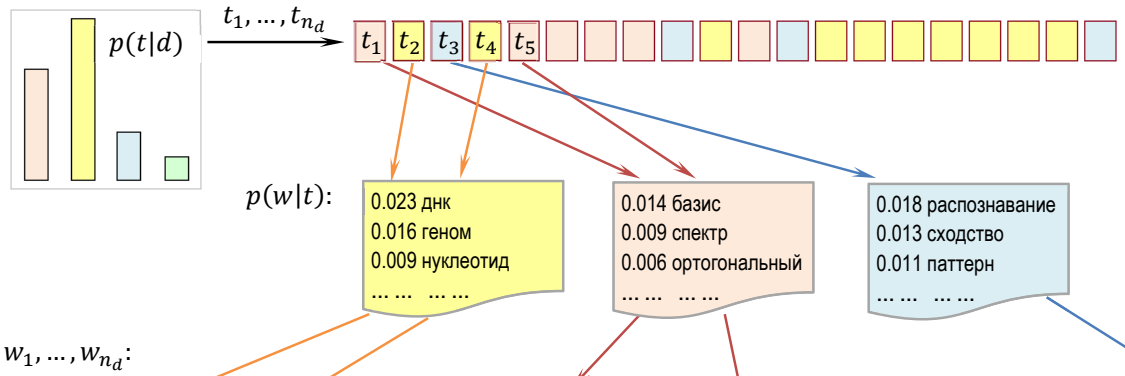
Тематическая модель для разведочного поиска должна быть...

1. **Интерпретируемая**: каждая тема понятна людям
2. **Иерархическая**: систематизация областей знания
3. **Темпоральная**: обнаружение и прослеживание тем
4. **Мультиграммная**: выделение тематичных словосочетаний
5. **Мультимодальная**: авторы, связи, тэги, пользователи,...
6. **Мультиязычная**: кросс- и много-языковой поиск
7. **Разреженная**: сокращение поискового индекса
8. **Сегментирующая**: выделение тем внутри документа
9. **Обучаемая**: учёт обратной связи с пользователями
10. **Создающая и именующая темы** автоматически
11. **Онлайновая**: обрабатывающая коллекцию за 1 проход
12. **Параллельная, распределённая**: для больших коллекций

Вероятностная модель порождения текстовой коллекции

объясняет появление терминов w в документах d темами 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).

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

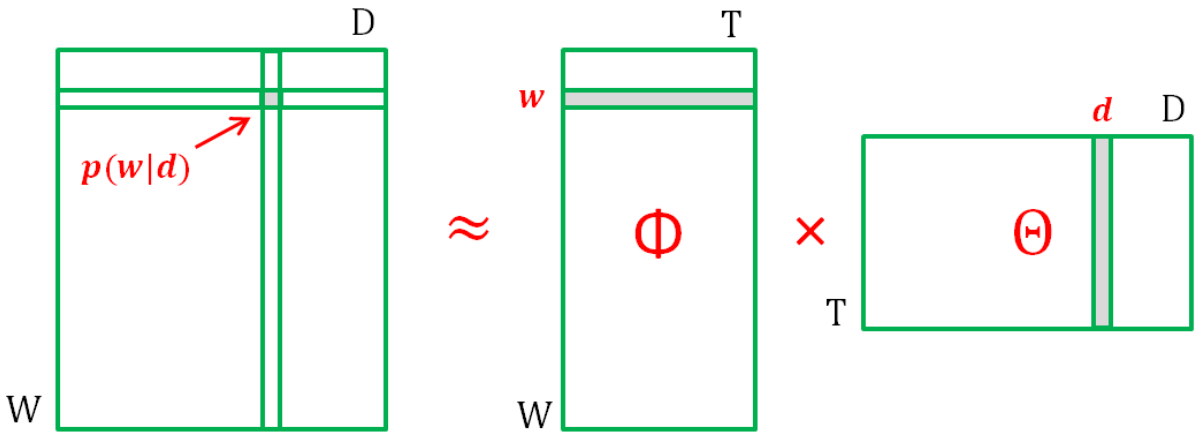
Дано: коллекция текстовых документов

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

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),

Задача стохастического матричного разложения является *некорректно поставленной* — её решение не единственно:

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

для невырожденных $S_{T\times T}$ таких, что Φ', Θ' — стохастические.

Регуляризация — дополнительные ограничения на Φ, Θ.

Основные вехи развития тематического моделирования

- PLSA (1999) решает нерегуляризованную задачу
- LDA (2003) регуляризатор Дирихле, самая известная модель
- ATM (2004) авторы документов
- TOT (2006) метки времени документов
- HDP (2006) определение числа тем
- TNG (2007) группирование слов в мультиграммы
- CTM (2007) корреляции между темами
- NetPLSA (2008) граф связей между документами
- ML-LDA (2009) многоязычные параллельные тексты
- ssLDA (2012) частичное обучение
- Dependency-LDA (2012) категоризация документов
- mLDA (2013) метаданные с тремя и более модальностями
- BitermTM (2013) анализ коротких документов
- WNTM (2014) локальные контексты слов

Байесовская регуляризация — доминирующий подход в BTM

Байесовский вывод — основа LDA и последующих моделей:

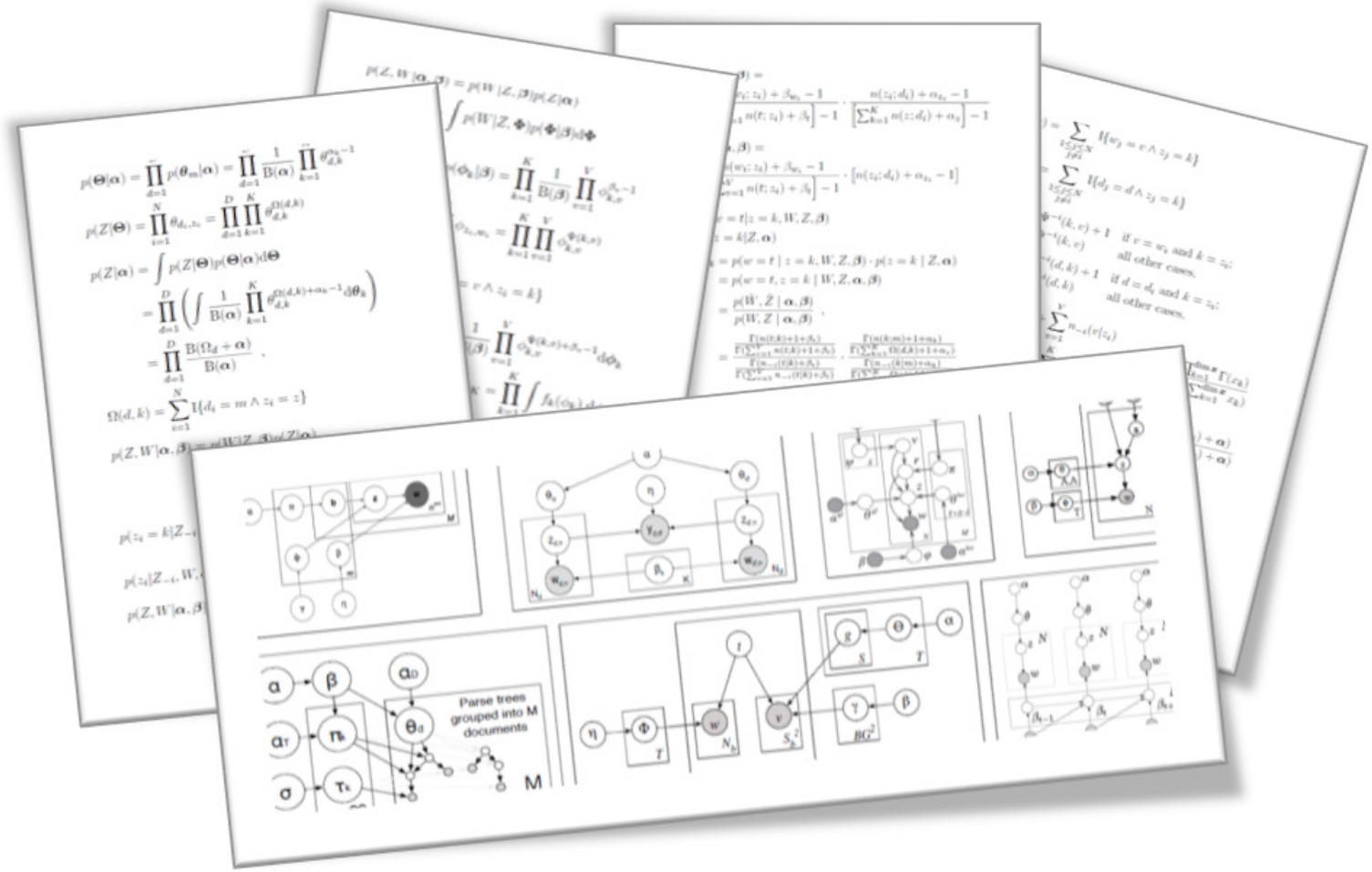
$$\text{Prior}(\Phi, \Theta) + \text{Data} \quad \rightarrow \quad \text{Posterior}(\Phi, \Theta).$$

Проблемы:

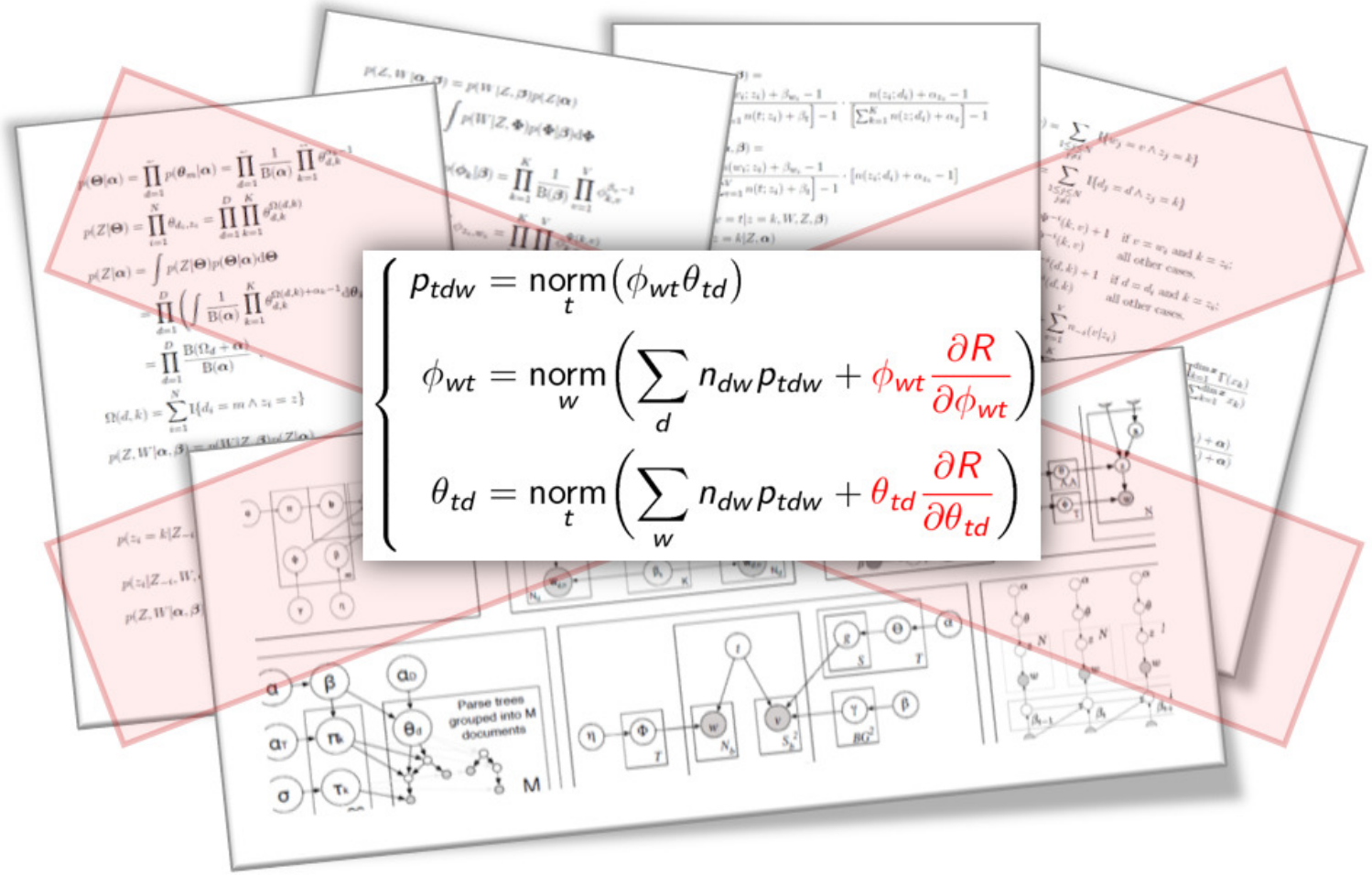
- Задача избыточно сложная: в тематическом моделировании нужны лишь значения Φ, Θ, а не их распределения
- Используемые $\text{Prior}(\Phi, \Theta)$ удобны для математических выкладок, но лингвистически не обоснованы
- Обучение каждой модели — уникальная задача
- Трудно унифицировать и комбинировать модели
- Плоская нотация не способствует пониманию моделей

Rob Zinkov. Stop using Plate Notation.
http://zinkov.com/posts/2013-07-28-stop-using-plates

Байесовское обучение — доминирующий подход в BTM



ARTM — альтернатива байесовскому обучению



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

ARTM — Аддитивная Регуляризация Тематических Моделей

Максимизация log правдоподобия с регуляризатором R:

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

EM-алгоритм: метод простой итерации для системы уравнений

$$\text{E-шаг:} \quad p_{tdw} = \underset{t\in T}{\text{norm}}\left(\phi_{wt}\theta_{td}\right)$$

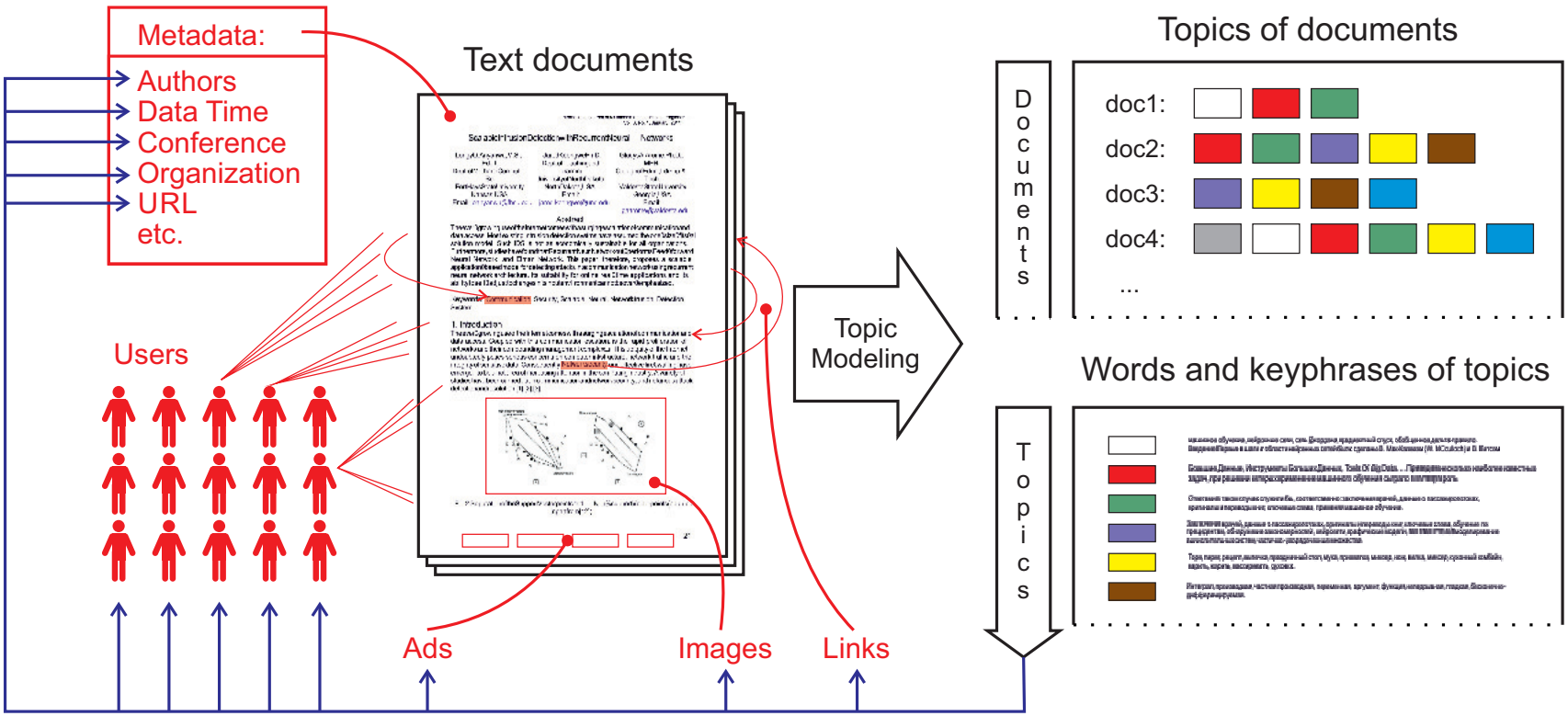
$$\text{M-шаг:} \quad \begin{cases} \phi_{wt} = \underset{w\in W}{\text{norm}}\left(\sum_{d\in D} n_{dw} p_{tdw} + \phi_{wt}\frac{\partial R}{\partial \phi_{wt}} \right) \\ \theta_{td} = \underset{t\in T}{\text{norm}}\left(\sum_{w\in d} n_{dw} p_{tdw} + \theta_{td}\frac{\partial R}{\partial \theta_{td}} \right) \end{cases}$$

Модель PLSA: $R(\Phi,\Theta) = 0$

Модель LDA: $R(\Phi,\Theta) = \sum_{t,w} \beta_w \ln \phi_{wt} + \sum_{d,t} \alpha_t \ln \theta_{td}$

Обобщение ARTM на мультимодальные задачи

Темы определяют распределения не только терминов $p(w|t)$,
но и других *модальностей*: $p(\text{автор}|t)$, $p(\text{время}|t)$, $p(\text{ссылка}|t)$,
$p(\text{баннер}|t)$, $p(\text{элемент\_изображения}|t)$, $p(\text{пользователь}|\text{t}), \ldots$



Обобщение ARTM на мультимодальные задачи

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

$W = W^1 \sqcup \cdots \sqcup W^M$ — объединённый словарь всех модальностей

Максимизация суммы log правдоподобий с регуляризацией:

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

EM-алгоритм: метод простой итерации для системы уравнений

$$\text{E-шаг:} \quad p_{tdw} = \underset{t \in T}{\text{norm}}\left(\phi_{wt} \theta_{td}\right)$$

$$\text{M-шаг:} \quad \begin{cases} \phi_{wt} = \underset{w \in W^m}{\text{norm}}\left(\sum_{d \in D} \lambda_{m(w)} n_{dw} p_{tdw} + \phi_{wt} \frac{\partial R}{\partial \phi_{wt}} \right) \\ \theta_{td} = \underset{t \in T}{\text{norm}}\left(\sum_{w \in d} \lambda_{m(w)} n_{dw} p_{tdw} + \theta_{td} \frac{\partial R}{\partial \theta_{td}} \right) \end{cases}$$

BigARTM: библиотека тематического моделирования

Ключевые возможности:

- Онлайновый 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: унификация разработки тематических моделей

Этапы моделирования	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	

-- *нестандартизуемые этапы, уникальная разработка для каждой задачи*

-- *стандартизуемые этапы*

Разработка тематических моделей в среде IPython Notebook

http://nbviewer.ipython.org/github/bigartm/bigartm-book/tree/master/

Коллекция:

Воспользуемся небольшой коллекцией 'kos', доступной в репозитории UCI https://archive.ics.uci.edu/ml/machine-learning-databases/bag-of-words/. Параметры коллекции следующие:

- 3430 документов;
- 6906 слов в словаре;
- 467714 слов в коллекции.

Для начала подключим все необходимые модули (убедитесь, что путь к Python API BigARTM находится в вашей переменной PATH):

In [1]:
```
%matplotlib inline
import glob
import matplotlib.pyplot as plt

import artm
```

Прежде всего необходимо подготовить входные данные. BigARTM имеет собственный формат документов для обработки, называемый батчами. В библиотеки присутствуют средства по созданию батчей из файлов Bag-Of-Words в форматах UCI и Vowpal Wabbit (подробности можно найти в http://docs.bigartm.org/en/latest/formats.html).

В Python API, по аналогии с алгоритмами из scikit-learn, входные данные представлены одним классом BatchVectorizer. Объект этого класса принимает на вход батчи или файлы с Bag-Of-Words и подаётся на вход всем методам. В случае, если входные данные не являются батчами, он создаёт их и сохраняет на диск для последующего быстрого использования.

Итак, создадим объект BatchVectorizer:

In [2]:
```
batch_vectorizer = None
if len(glob.glob('kos' + "/*.batch")) < 1:
    batch_vectorizer = artm.BatchVectorizer(data_path='', data_format='bow
_uci', collection_name='kos', target_folder='kos')
else:
    batch_vectorizer = artm.BatchVectorizer(data_path='kos', data_format='
batches')
```

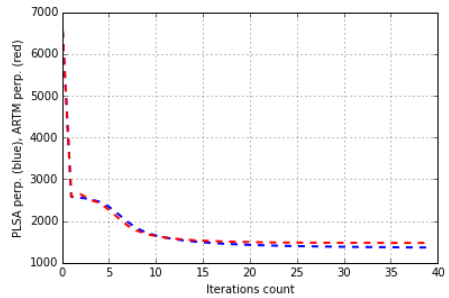
ARTM --- это класс, представляющий собой Python API BigARTM, и позволяющий использовать практически все возможности библиотеки в стиле scikit-learn. Создадим две тематические модели для нашего эксперимента. Наиболее важным параметром модели является число тем. Опционально можно указать списки регуляризаторов и функционалов качества, которые следует использовать для данной модели. Если этого не сделать, то регуляризаторы и функционалы всегда можно добавить позднее. Обратите внимание, что каждая модель задаёт

Продолжим обучение моделей, инициировав 25 проходов по коллекции, после чего снова посмотрим на значения функционалов качества:

In [11]:
```
model_plsa.fit_offline(batch_vectorizer=batch_vectorizer, num_collection_p
asses=25, num_document_passes=1)
model_artm.fit_offline(batch_vectorizer=batch_vectorizer, num_collection_p
asses=25, num_document_passes=1)
```

In [12]:
```
print_measures(model_plsa, model_artm)
```

```
Sparsity Phi: 0.332 (PLSA) vs. 0.740 (ARTM)
Sparsity Theta: 0.082 (PLSA) vs. 0.602 (ARTM)
Kernel contrast: 0.530 (PLSA) vs. 0.568 (ARTM)
Kernel purity: 0.396 (PLSA) vs. 0.531 (ARTM)
Perplexity: 1365.804 (PLSA) vs. 1475.455 (ARTM)
```



Кроме того, для наглядности построим графики изменения разреженностей матриц по итерациям:

In [13]:
```
plt.plot(xrange(model_plsa.num_phi_updates), model_plsa.score_tracker['Spa
rsityPhiScore'].value, 'b--',
                 xrange(model_artm.num_phi_updates), model_artm.score_trac
ker['SparsityPhiScore'].value, 'r--', linewidth=2)
plt.xlabel('Iterations count')
plt.ylabel('PLSA Phi sp. (blue), ARTM Phi sp. (red)')
plt.grid(True)
plt.show()

plt.plot(xrange(model_plsa.num_phi_updates), model_plsa.score_tracker['Spa
rsityThetaScore'].value, 'b--',
                 xrange(model_artm.num_phi_updates), model_artm.score_trac
ker['SparsityThetaScore'].value, 'r--', linewidth=2)
```

Тесты производительности

- 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* = перплексия на тестовой выборке документов

Мониторинг межнациональных отношений в соцсети

Дано: посты социальных сетей.

Найти: темы, связанные с межнациональными отношениями.

Регуляризаторы:

- частичное обучение тем по словарю этнонимов
- учёт модальностей геолокаций и времени
- повышение различности тем
- выделение фоновых тем общей лексики

M.Apishev, S.Koltcov, O.Koltsova, S.Nikolenko, K.Vorontsov. Mining Ethnic Content Online with Additively Regularized Topic Models. Computación y Sistemas, 2016.

Тематизация новостных потоков для медиапланирования

Дано: новости от многих источников СМИ.

Найти: иерархию и динамику развития тем;
оценки сходства тематических спектров новостных источников.

Регуляризаторы:

- связывание уровней иерархии
- учёт модальностей времени и источников
- повышение различности тем
- выделение фоновых тем общей лексики

Сценарный анализ записей разговоров контакт-центра

Дано: текстовые записи разговоров контакт-центра.

Найти: тематическую сегментацию каждого разговора; отклонения от инструкций в работе операторов; автоматизировать переключения и подсказки оператору.

Регуляризаторы:

- частичное обучение тем по текстам инструкций
- определение границ тематических сегментов
- повышение различности тем
- выделение фоновых тем общей лексики

Разведочный тематический поиск

Дано: тексты коллективного блога (Хабрахабр)

Найти: алгоритм ранжирования по тематической близости и оценить качество тематического поиска.

Регуляризаторы:

- разреживание $p(t|d)$ для сокращения поискового индекса
- учёт модальностей авторов, комментаторов, тегов, хабов
- повышение различности тем
- выделение фоновых тем общей лексики

А.О.Янина, К.В.Воронцов. Мультимодальные тематические модели для разведочного поиска в коллективном блоге. JMLDA, 2016.

Скрининговая диагностика по электрокардиограмме

Дано: дискретизированные ЭКГ-сигналы + диагнозы (данные о вариабельности сердечного ритма, закодированные в виде символьной последовательности)

Найти: темы — диагностические эталоны заболеваний; алгоритм оценивания риска болезни.

Регуляризаторы:

- учёт диагнозов в обучающей выборке
- повышение различности тем
- выделение фоновых тем

Анализ данных о транзакциях клиентов банка

Дано:

D — множество клиентов (15 000)

W — категории = MCC-коды (Merchant Category Code) (328)

n_{dw} — сумма транзакций клиента d по категории w

Найти: темы — типы экономического поведения (потребления)

$\phi_{wt} = p(w|t)$ — структура потребления для темы t

$\theta_{td} = p(t|d)$ — типы потребления клиента d

Регуляризаторы:

- повышение различности тем
- разреживание $p(t|d)$
- учёт модальностей времени, типа транзакции, терминала
- обучение прогнозов объёма трат по метрике RMSLE_1

Интерпретация тематической модели

Прогноз объёма трат клиента d по категории w

$$n_{dw} = n_d \sum_{t \in T} p(w|t)p(t|d),$$

n_d — объём клиента d по всем категориям или его прогноз.

Проблема несбалансированности клиентов по объёмам:
приоритет получают «богатые» клиенты с наибольшими n_d

$$\sum_{d \in D} \sum_{w \in d} n_{dw} \ln \sum_{t \in T} \phi_{wt}\theta_{td} \to \max_{\Phi,\Theta}$$

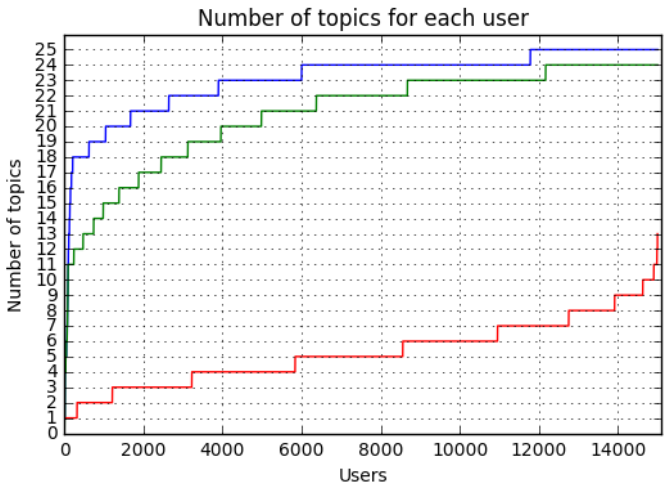
Два способа решения (первый сохраняет аддитивность объёмов):

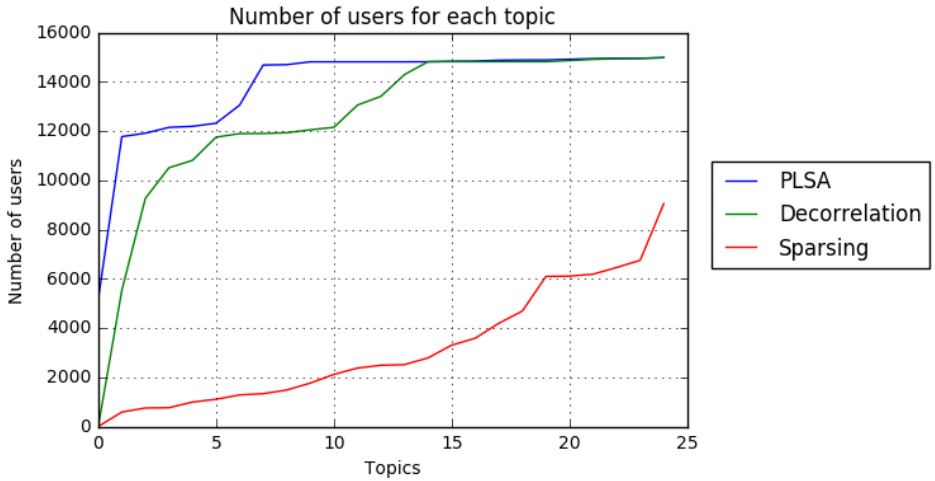
$$\sum_{d \in D} \sum_{w \in d} \frac{1}{n_d} n_{dw} \ln \sum_{t \in T} \phi_{wt}\theta_{td} \to \max_{\Phi,\Theta};$$

$$\sum_{d \in D} \sum_{w \in d} \log n_{dw} \ln \sum_{t \in T} \phi_{wt}\theta_{td} \to \max_{\Phi,\Theta}$$

Построение модели ARTM, 25 тем

- 30 итераций PLSA — без регуляризаторов
- 10 итераций — повышение различности тем
- 10 итераций — разреживание $p(t|d)$





Декоррелирование Φ и разреживание Θ определяют минимальное число типов экономического поведения каждого клиента, достаточное для описания его расходов.

Пользуюсь картой только чтобы снять наличные

ϕ_{wt},%	MCC-код (категория расходов)
72	Финансовые институты — снятие наличности вручную
27	Финансовые институты — снятие наличности автоматически
0.23	Денежные переводы MasterCard MoneySend
0.1	Денежные переводы
0.012	Финансовые институты — снятие наличности вручную
0.0055	Легковой и грузовой транспорт: продажа, сервис, ремонт, лизинг
0.0027	Магазины игрушек

Наличные + авто, спорт, компьютеры

ϕ_{wt},%	MCC-код (категория расходов)
55	Финансовые институты — снятие наличности автоматически
44	Денежные переводы
0.111	Станции техобслуживания
0.105	Автозапчасти и аксессуары
0.094	Компьютерная сеть/информационные услуги
0.043	Спортивная одежда, одежда для верховой езды и езды на мотоцикле
0.024	Финансовые институты — снятие наличности вручную
0.020	СТО общего назначения
0.018	Горючее топливо — уголь, нефть, разжиженный бензин, дрова
0.015	Магазины мужской и женской одежды
0.015	Финансовые институты — снятие наличности вручную
0.013	Магазины спорттоваров
0.012	Садовые принадлежности (в том числе для ухода за газонами) в розницу
0.011	Паркинги и гаражи
0.011	Бакалейные магазины, супермаркеты
0.010	Различные магазины одежды и аксессуаров

Цивилизованный потребитель: разные магазины, связь, авто

ϕ_{wt},%	MCC-код (категория расходов)
27	Станции техобслуживания
20	Различные продовольственные магазины, рынки, полуфабрикаты
15	Звонки с использованием телефонов, считывающих магнитную ленту
12	Финансовые институты — снятие наличности автоматически
4.7	Горючее топливо — уголь, нефть, разжиженный бензин, дрова
4.1	Универсальные магазины
3.4	Автозапчасти и аксессуары
1.4	Аптеки
1.2	Магазины с продажей спиртных напитков на вынос
1.1	Бакалейные магазины, супермаркеты
0.57	Автошины
0.37	Прямой маркетинг — торговля через каталог
0.35	Товары для дома
0.33	Универмаги
0.32	Плавательные бассейны — распродажа
0.21	Места общественного питания, рестораны

Всего 24 категории с $\phi_{wt} > 0.1\%$; 61 категория с $\phi_{wt} > 0.01\%$

Продвинутые мамки

ϕ_{wt},%	MCC-код (категория расходов)
56	Бакалейные магазины, супермаркеты
8.6	Финансовые институты — снятие наличности автоматически
5.4	Аптеки
4.0	Звонки с использованием телефонов, считывающих магнитную ленту
2.2	Рестораны, закусочные
1.8	Обувные магазины
1.5	Различные продовольственные магазины — рынки, полуфабрикаты
1.4	Магазины спорттоваров
1.4	Детская одежда, включая одежду для самых маленьких
1.3	Магазины игрушек
1.3	Места общественного питания, рестораны
1.1	Магазины мужской и женской одежды
1.1	Магазины с продажей спиртных напитков на вынос
1.1	Магазины косметики
1.0	Садовые принадлежности в розницу
0.73	Одежда для всей семьи

Всего 41 категория с $\phi_{wt} > 0.1\%$; 95 категорий с $\phi_{wt} > 0.01\%$

Бизнес-леди: забыла про наличку — всё по карте

ϕ_{wt},%	MCC-код (категория расходов)
12	Магазины мужской и женской одежды
7.3	Оборудование, мебель и бытовые принадлежности
7.0	Места общественного питания, рестораны
5.6	Магазины по продаже часов, ювелирных изделий и изделий из серебра
5.3	Обувные магазины
4.7	Магазины косметики
4.6	Одежда для всей семьи
3.8	Универмаги
3.2	Готовая женская одежда
2.8	Практикующие врачи, медицинские услуги
1.8	Прямой маркетинг — торговля через каталог
1.5	Салоны красоты и парикмахерские
1.3	Детская одежда, включая одежду для самых маленьких
1.3	Аптеки
1.0	Изготовление и продажа меховых изделий
1.0	Центры здоровья

Всего 70 категорий с $\phi_{wt} > 0.1\%$; 134 категории с $\phi_{wt} > 0.01\%$

Продвинутый активный потребитель всего, и по карте

ϕ_{wt},%	MCC-код (категория расходов)
20	Финансовые институты — снятие наличности вручную
15	Универсальные магазины
13	Туристические агентства и организаторы экскурсий
11	Автозапчасти и аксессуары
8.8	Коммунальные услуги — электричество, газ, санитария, вода
4.2	Веломагазины — продажа и обслуживание
3.7	СТО общего назначения
0.9	Услуги курьера — по воздуху и на земле, агентство по отправке грузов
0.8	Рекламные услуги
0.7	Компьютеры, периферия, программное обеспечение
0.5	Образовательные услуги
0.4	Бакалейные магазины, супермаркеты
0.4	Практикующие врачи, медицинские услуги
0.3	Продажа мотоциклов
0.3	Оборудование, мебель и бытовые принадлежности
0.2	Автошины

Всего 35 категорий с $\phi_{wt} > 0.1\%$; 93 категории с $\phi_{wt} > 0.01\%$

Бизнес-класс: авиа, отели, казино, рестораны, ценные бумаги

ϕ_{wt},%	MCC-код (категория расходов)
28	Авиалинии, авиакомпании
19	Финансовые институты — торговля и услуги
9.5	Отели, мотели, базы отдыха, сервисы бронирования
8.6	Транзакции по азартным играм (плюс)
5.2	Финансовые институты — торговля и услуги
3.2	Места общественного питания, рестораны
3.1	Не-финансовые институты: ин.валюта, переводы, дорожн.чеки, квази-кэш
2.2	Пассажирские железнодорожные перевозки
1.7	Бизнес-сервис
1.4	Жилье — отели, мотели, курорты
1.3	Галереи/учреждения видеоигр
1.3	Транзакции по азартным играм (минус)
0.6	Ценные бумаги: брокеры/дилеры
0.5	Туристические агентства и организаторы экскурсий
0.3	Лимузины и такси
0.3	Беспошлинные магазины Duty Free

Всего 50 категорий с $\phi_{wt} > 0.1\%$; 103 категории с $\phi_{wt} > 0.01\%$

Провинциальный малый бизнес

ϕ_{wt},%	MCC-код (категория расходов)
27	Финансовые институты — снятие наличности автоматически
8.5	Лесо- и строительный материал
8.4	Бытовое оборудование
6.6	Плавательные бассейны — распродажа
5.5	Продажа электронного оборудования
4.1	Бакалейные магазины, супермаркеты
3.3	Универсальные магазины
3.0	Садовые принадлежности в розницу
2.6	Телекоммуникационное оборудование, включая продажу телефонов
2.4	Легковой и грузовой транспорт: продажа, сервис, ремонт, лизинг
2.2	Товары для дома
2.1	Пассажирские железнодорожные перевозки
1.5	Оборудование, мебель и бытовые принадлежности
1.3	Скобяные товары в розницу
1.2	Магазины спорттоваров
1.1	Аптеки

Всего 54 категории с $\phi_{wt} > 0.1\%$; 104 категории с $\phi_{wt} > 0.01\%$

Резюме

- Когда нужна тематическая модель, все берут LDA, который может давать неустойчивые результаты, и пользуются пост-обработкой тем, найденных LDA
- ARTM позволяет оптимизировать темы под задачу, строить модели с заданными свойствами, комбинируя различные критерии и источники данных
- Новое применение тематического моделирования — анализ транзакционных данных банка

http://bigartm.org

Join BigARTM community!

Литература

- *Воронцов К. В.* Аддитивная регуляризация тематических моделей коллекций текстовых документов. Доклады РАН. 2014.
- *K.Vorontsov, A.Potapenko.* Tutorial on probabilistic topic modeling: Additive regularization for stochastic matrix factorization. AIST 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. (в печати)
- *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. (в печати)
- *А.О.Янина, К.В.Воронцов.* Мультимодальные тематические модели для разведочного поиска в коллективном блоге. ИОИ 2016. (в печати)