МАШИННЫЙ ИНТЕЛЛЕКТ И УМНЫЙ ИНФОРМАЦИОННЫЙ ПОИСК

К. В. ВОРОНЦОВ,

Д.Ф.-М.Н., ПРОФЕССОР РАН,
ЗАВЕДУЮЩИЙ ЛАБОРАТОРИЕЙ МАШИННОГО ИНТЕЛЛЕКТА МФТИ

МФТИ

Я ПРОФИ ЗИМНЯЯ ШКОЛА ПО ИСКУССТВЕННОМУ ИНТЕЛЛЕКТУ

Содержание

Машинный интеллект — новый двигатель прогресса

«Четвёртая технологическая революция строится на вездесущем и мобильном Интернете, искусственном интеллекте и машинном обучении» (2016)

Клаус Мартин Шваб, президент Всемирного экономического форума

Стоит ли верить, что умные машины завоюют мир? Оставят нас без работы? Возможен ли общий искусственный интеллект? Опасно ли его самоосознание?

Бум искусственного интеллекта и нейронных сетей

1997 IBM Deep Blue обыграл чемпиона мира по шахматам

2005 Беспилотный автомобиль: DARPA Grand Challenge

2006 Google Translate – статистический машинный перевод

2011 40 лет DARPA CALO привели к созданию Apple Siri

2011 IBM Watson победил в ТВ-игре «Jeopardy!»

2011–2015 ImageNet: 25% $\rightarrow$ 3.5% ошибок против 5% у людей

2012 Google X Lab: распознавание видеокадров с котами

2014 Facebook DeepFace распознаёт лица с точностью 97%

2016 DeepMind, OpenAI: динамическое обучение играм Atari

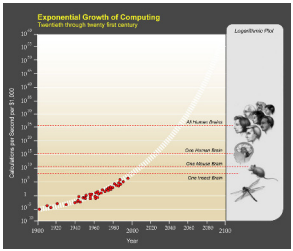
2016 Google DeepMind обыграл чемпиона мира по игре го

2017 OpenAI обыграл чемпиона мира по компьютерной игре Dota 2

Три предпосылки этого бума

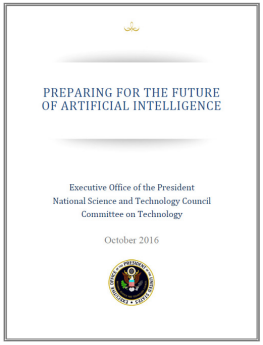
Три перехода количества в качество в нейронных сетях:

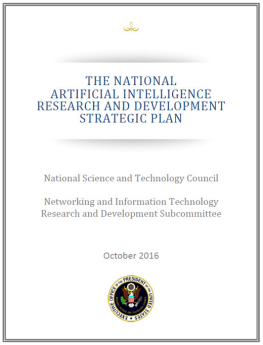
1. **Достижения микроэлектроники**
   — процессоры, память, графические карты
   — рост вычислительных мощностей по закону Мура
   — экстраполяция: $80 \cdot 10^9$ нейронов в 2035–2050 гг.
2. **Повсеместность и доступность IT-технологий**
   — накопление больших выборок данных
   — краудсорсинг (пример — ImageNet)
3. **Развитие методов машинного обучения**
   — rectified linear unit, ReLU (V.Nair & G.Hinton, 2010)
   — быстрые SGD алгоритмы: AdaDelta (Kingma & Ba 2014)
   — dropout (G.Hinton, 2012), регуляризации

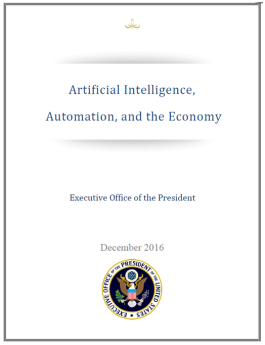


Ray Kurzweil. The singularity is near: When humans transcend biology. 2006.

Отчёты Белого Дома США, май–октябрь 2016







«Nations with the strongest presence in AI R&D will establish leading positions in the automation of the future»

Preparing for the Future of Artificial Intelligence. NSTC. 2016.

Отчёты Белого Дома США, октябрь 2016

- Сокращение издержек и повышение производительности труда
- Автоматизация банковских и финансовых услуг (FinTech)
- Автоматизация юридических услуг (LegalTech)
- Автоматизация посредничества и распределённая экономика
- Оптимизация логистики, энергетических и транспортных сетей
- Роботизация производства и автономный транспорт
- Сенсорные сети, мониторинг сельского хозяйства
- Персональная медицина, улучшение клинических практик
- Персонализация образования (EdTech) и социальная инженерия
- Автономные системы вооружений

Preparing for the Future of Artificial Intelligence. NSTC. 2016.

Отчёты Белого Дома США. Некоторые из 23 рекомендаций

1. Государственным и коммерческим организациям: развивать партнёрство с научными коллективами для эффективного использования данных
2. Развивать стандарты открытых данных для привлечения научного сообщества к решению задач
8. Развивать системы управления беспилотным транспортом
11. Вести постоянный мониторинг исследований ИИ в мире
13. Поддерживать фундаментальные исследования по ИИ
14. Развивать образовательные программы по ИИ и курсы повышения квалификации для прикладников
20. Развивать международную кооперацию по ИИ
22. Учитывать взаимовлияние ИИ и кибербезопасности

Preparing for the Future of Artificial Intelligence. NSTC. 2016.

Особенности реальных данных

В прикладных задачах данные бывают...

- разнородные (признаки измерены в разных шкалах)
- неполные (признаки измерены не все, имеются пропуски)
- неточные (признаки измерены с погрешностями)
- противоречивые (объекты одинаковые, ответы разные)
- избыточные (сверхбольшие, не помещаются в память)
- недостаточные (объектов меньше, чем признаков)
- неструктурированные (нет признаковых описаний)
- «грязные» (ошибочные, грубо не соответствующие истине)

Для всех этих случаев известны специальные подходы... кроме грязных данных!

Особенности реальных проектов

Проблема №1: некомпетентный исполнитель

- не готов к преодолению сложностей реальных задач

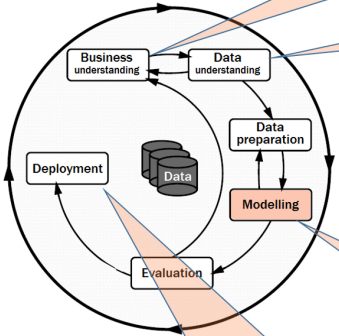
Проблема №2: некомпетентный заказчик

- ждёт чуда от искусственного интеллекта и больших данных
- не имеет численных критериев качества (KPI)
- не заботится о чистоте данных
- не готов пилотировать новые технологии
- не отличает простые задачи от сложных

Необходимость инвестиционных и образовательных проектов

- цифровая трансформация бизнес-процессов
- введение контроля качества данных
- кооперация бизнеса, науки и образования
- проведение конкурсов на открытых данных

Факторы риска и точки приложения силы



Несовершенство бизнес-процессов сбора данных

Системный анализ и реорганизация процессов

Грязные данные

Конкурсы на открытых данных

Нехватка ресурсов на исследования

Нехватка квалификации

Кооперация науки, образования и бизнеса

Нехватка кадров

Создание новых методов для новых задач

Нестандартные задачи и модели

Платформы адаптивного обучения моделей

Задержки между получением данных, моделированием и внедрением

Открытые данные. Открытый код. Открытая наука

- **Выгоды открытых данных**
  - *для индустрии:* бенчмаркинг, стандартизация, привлечение внимания
  - *для компаний:* подбор исполнителей, сокращение издержек и рисков
  - *для университетов:* интеграция практических задач в учебный процесс
  - *для исследователей:* проверка новых теорий и технологий в деле
  - *для студентов:* получение опыта, наработка портфолио
- **Выгоды открытого кода**
  - снижение издержек, ускорение разработки и внедрения
  - координация усилий исследователей и разработчиков
  - снижение технологических барьеров для выхода на рынок
- **Конкурсы анализа данных**
  - www.NetflixPrize.com – первый крупный конкурс, \$1 млн. (2006-2009)
  - www.kaggle.com – самая известная платформа
  - DataRing.ru – отечественная конкурсная платформа

Кооперация бизнеса, науки и образования

Проблемы: различия в целях, «некомпетентность», дефицит доверия

Опыт кафедры «Интеллектуальные системы» МФТИ

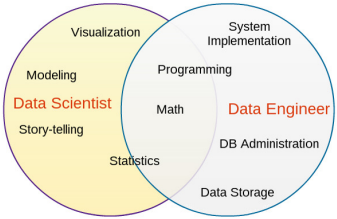
- Практикум В.В.Стрижова (страница на www.MachineLearning.ru)
  — более 700 индивидуальных студенческих проектов за 12 лет
- Начало сотрудничества — пилотный проект в рамках практикума

Шаги долгосрочного сотрудничества

- НИР/ОКР для университетской лаборатории
- формирование постоянной проектной группы, стажерская программа
- формирование образовательных курсов/модулей по решенным задачам
- открытие собственной лаборатории или кафедры
- тесная кооперация с собственным исследовательским отделом

Рынок труда в области анализа данных

- Инженер по данным (Data Engineer)
  - Понимает бизнес-процессы, порождающие данные
  - Работает с сырыми данными в различных форматах
  - Визуализирует, понимает, очищает, готовит данные
- Исследователь данных (Data Scientist)
  - Моделирует, строит признаки (feature engineering)
  - Выбирает модели и методы, оценивает решения
  - Ходит по кругу CRISP-DM
- Менеджер проектов по анализу данных
  - Организует бизнес-процессы сбора и очистки данных
  - Видит бизнес задачи и формализует их в терминах «Дано-Найти-Критерий»
  - Организует открытые конкурсы и пилотные проекты
  - Адекватно оценивает сложность задач и трудозатраты

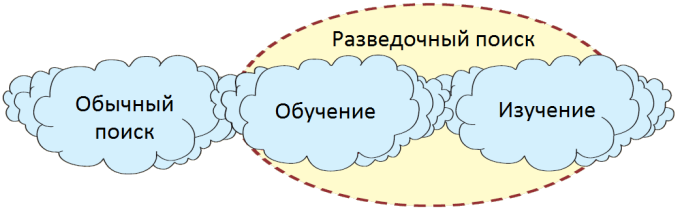


Исследования лаборатории Машинного Интеллекта МФТИ

- Анализ текстов и разведочный информационный поиск
  - Тематическое моделирование и поиск близких по смыслу документов
  - Тематизация диалогов контактных центров, классификация интентов
  - Проект TopicNet: иерархическое тематическое моделирование
- Анализ транзакционных данных
  - Профилирование потребительского поведения розничных клиентов банка
  - Выявление видов экономической деятельности корпоративных клиентов банка
- Анализ изображений и видео
  - Распознавание текстов на сканах/фотоснимках документов
- Анализ и прогнозирование временных рядов, интернет вещей
  - Классификация физической активности человека по данным датчиков
  - Анализ биомедицинских сигналов (ЭКГ, БКГ, ЭЭГ, ЭКоГ)

Концепция разведочного поиска (exploratory search)

- пользователь может не знать ключевых терминов
- запросом может быть текст произвольной длины
- информационная потребность — систематизация знаний



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

Что такое «тема» в коллекции текстовых документов?

Выделение тем — первый шаг к пониманию смысла текста

- *тема* — специальная терминология предметной области
- *тема* — набор часто совместно встречающихся слов и словосочетаний

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.

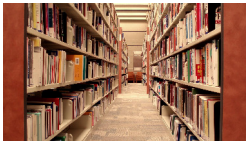
О какой теме t думал автор, когда писал термин w в документе d?

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

Приложения тематического моделирования

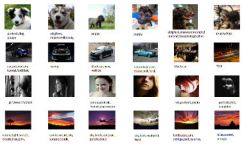
Тематическое моделирование — «мягкая кластеризация» коллекции текстов

разведочный поиск в электронных библиотеках

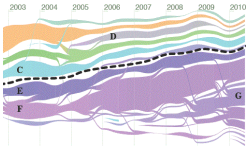


персонализированный поиск в соцсетях

мультимодальный поиск текстов и изображений

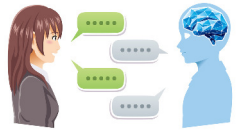


детектирование и трекинг новостных сюжетов



навигация по большим текстовым коллекциям

управлением диалогом в разговорном интеллекте



Пример. Мультиязычная модель Википедии. Интерпретируемость тем.

216K русско-английских пар статей. Первые 10 слов и их вероятности в теме, %:

Тема №68				Тема №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

Асессор оценил 396 тем из 400 как хорошо интерпретируемые.

K.Vorontsov, O.Frei, M.Apishev, P.Romov, M.Suvorova. BigARTM: Open Source Library for Regularized Multimodal Topic Modeling of Large Collections. AIST-2015.

Пример. Мультиязычная модель Википедии. Интерпретируемость тем.

216K русско-английских пар статей. Первые 10 слов и их вероятности в теме, %:

Тема №88				Тема №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

Асессор оценил 396 тем из 400 как хорошо интерпретируемые.

K.Vorontsov, O.Frei, M.Apishev, P.Romov, M.Suvorova. BigARTM: Open Source Library for Regularized Multimodal Topic Modeling of Large Collections. AIST-2015.

Поиск тематически близких документов

$\theta_{tq} = p(t|q)$ — тематический профиль текста запроса q
$\theta_{td} = p(t|d)$ — тематические профили документов d из коллекции

Косинусная мера близости документа d и запроса q:

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

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

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

Две коллекции новостей про технологии

Habrahabr.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

TechCrunch

Методика оценивания качества разведочного поиска

Поисковый запрос

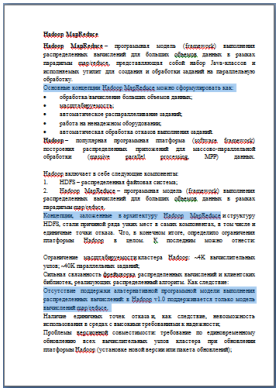
ключевые слова или фрагменты текста, одна страница A4

Поисковая выдача

документы, тематически близкие к документу-запросу

Два задания асессорам

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



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 A4):

Алгоритмы раскраски графов	Система IBM Watson
Рекомендательная система Netflix	3D-принтеры
Методики быстрого набора текста	CERN-кластер
Космические проекты Илона Маска	AB-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

Оценивание качества поиска

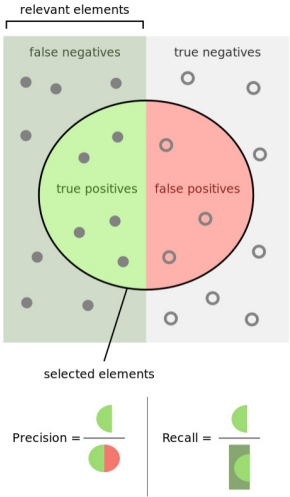
Precision — доля релевантных среди найденных
Recall — доля найденных среди релевантных

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

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

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

TP (true positive) — найденные релевантные
FP (false positive) — найденные нерелевантные
FN (false negative) — ненайденные релевантные



Какие модели поиска сравнивались

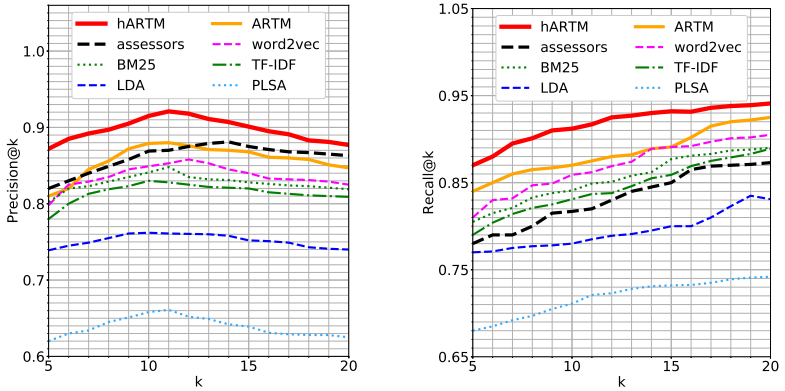
- **assessors:** результаты поиска, выполненного людьми (асессорами)
- **TF-IDF, BM25:** сравнение документов по векторам частот слов
- **word2vec:** нетематические векторные представления слов
- **PLSA:** Probabilistic Latent Semantic Analysis [T.Hofmann, 1999]
- **LDA:** Latent Dirichlet Allocation [D.Blei, A.Ng, M.Jordan, 2003]
- **ARTM:** тематическая модель с тремя регуляризаторами
- **hARTM:** двухуровневая иерархическая тематическая модель

Задачи регуляризаторов в ARTM и hARTM:

- сделать темы как можно более различными
- сделать профили $p(t|d)$ как можно более разреженными
- не допустить разреживания $p(w|t)$ до вырожденного состояния

Сравнение качества поиска с асессорами и простыми моделями

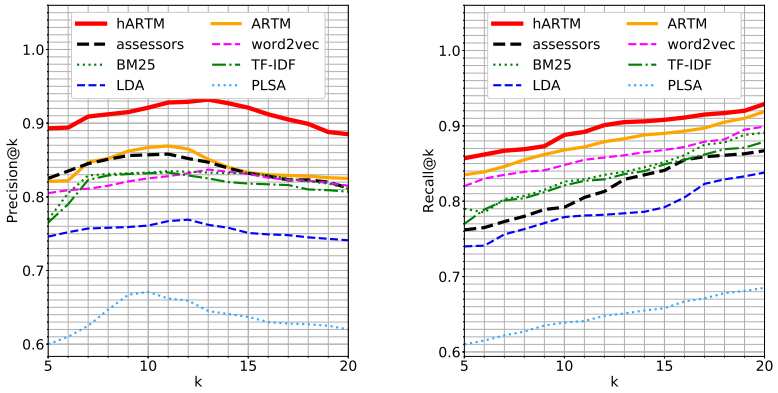
Точность и полнота по первым k позициям поисковой выдачи (Habrahabr.ru)



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

Сравнение качества поиска с асессорами и простыми моделями

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

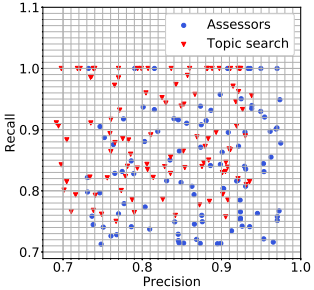


A.Ianina, K.Vorontsov. Multi-objective topic modeling for exploratory search in tech news. AINL, 2017.

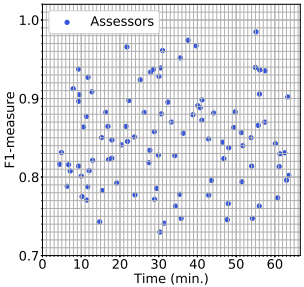
Результаты измерения точности и полноты по запросам

Точность, полнота и время поиска (100 запросов, 3 асессора на запрос, Habrahabr.ru)

точность и полнота поиска



время и F_1-мера (асессоры)



- среднее время обработки запроса асессором — 30 минут
- точность выше у асессоров, полнота — у поисковика

Влияние числа тем на качество поиска

Коллекция Habrahabr.ru

Используем 3 регуляризатора, 5 модальностей, меняем число тем

	acессоры	100	150	**200**	250	400
Prec@5	0.821	0.662	0.721	**0.810**	0.761	0.693
Prec@10	0.869	0.761	0.812	**0.879**	0.825	0.673
Prec@15	0.875	0.733	0.795	**0.868**	0.791	0.651
Prec@20	0.863	0.724	0.795	**0.847**	0.792	0.642
Recall@5	0.780	0.732	0.807	**0.840**	0.821	0.721
Recall@10	0.817	0.771	0.843	**0.870**	0.851	0.751
Recall@15	0.850	0.824	**0.895**	0.891	0.871	0.773
Recall@20	0.873	0.857	0.905	**0.925**	0.892	0.771

- Существует оптимальное по критерию качества поиска число тем

Влияние числа тем на качество поиска

Коллекция TechCrunch.com

Используем 3 регуляризатора, 4 модальности, меняем число тем

	acecсоры	350	400	450	**475**	500
Prec@5	0.822	0.653	0.725	0.752	**0.819**	0.777
Prec@10	0.851	0.663	0.732	0.762	**0.867**	0.811
Prec@15	0.835	0.682	0.743	0.787	**0.833**	0.793
Prec@20	0.813	0.650	0.743	0.773	**0.825**	0.793
Recall@5	0.762	0.731	0.762	0.793	**0.835**	0.817
Recall@10	0.792	0.763	0.793	0.812	**0.868**	0.855
Recall@15	0.835	0.782	0.807	0.855	**0.890**	0.882
Recall@20	0.867	0.792	0.823	0.862	**0.919**	0.903

- Оптимальное число тем существенно зависит от коллекции

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	**0.810**	0.652	0.775	0.779	**0.819**
Prec@10	0.653	0.776	0.812	**0.879**	0.679	0.787	0.819	**0.867**
Prec@15	0.642	0.765	0.792	**0.868**	0.669	0.773	0.798	**0.833**
Prec@20	0.643	0.759	0.783	**0.847**	0.673	0.777	0.792	**0.825**
Recall@5	0.692	0.784	0.805	**0.840**	0.673	0.812	0.812	**0.835**
Recall@10	0.714	0.814	0.834	**0.870**	0.685	0.821	0.845	**0.868**
Recall@15	0.725	0.835	0.867	**0.891**	0.712	0.859	0.869	**0.890**
Recall@20	0.735	0.862	0.891	**0.925**	0.723	0.882	0.895	**0.919**

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

Влияние сочетания модальностей на качество поиска

Коллекция Habrahabr.ru. Число тем $|T| = 200$. Модальности: Слова, Биграммы, Теги, Хабы, Комментаторы, Авторы.

	асессоры	С	К	СБ	СБТХ	все
Prec@5	0.821	0.612	0.549	0.654	0.737	**0.810**
Prec@10	0.869	0.635	0.568	0.701	0.752	**0.879**
Prec@15	0.875	0.625	0.532	0.685	0.682	**0.868**
Prec@20	0.863	0.616	0.533	0.682	0.687	**0.847**
Recall@5	0.780	0.722	0.636	0.797	0.827	**0.840**
Recall@10	0.817	0.744	0.648	0.812	0.875	**0.870**
Recall@15	0.850	0.778	0.677	0.842	0.893	**0.891**
Recall@20	0.873	0.803	0.685	0.852	0.898	**0.925**

- Наилучшее качество поиска — по всем модальностям
- Наиболее полезные модальности — слова и теги

Задача тематического моделирования

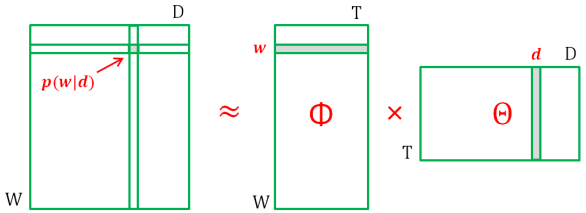
Дано: коллекция текстовых документов

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

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

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

Это задача стохастического матричного разложения:



PLSA — Probabilistic Latent Semantic Analysis [T.Hofmann, 1999]

Максимизация log-правдоподобия при $\phi_{wt} \geqslant 0$, $\theta_{td} \geqslant 0$, $\sum_w \phi_{wt} = 1$, $\sum_t \theta_{td} = 1$:

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

EM-алгоритм: метод простой итерации для системы уравнений

$$\begin{aligned}
&\text{E-шаг:} \quad p_{tdw} = p(t|d,w) = \operatorname*{norm}_{t\in T}\big(\phi_{wt}\theta_{td}\big) \\
&\text{M-шаг:} \quad \begin{cases} \phi_{wt} = \operatorname*{norm}_{w\in W}\Big(\sum\limits_{d\in D} n_{dw} p_{tdw}\Big) \\ \theta_{td} = \operatorname*{norm}_{t\in T}\Big(\sum\limits_{w\in d} n_{dw} p_{tdw}\Big) \end{cases}
\end{aligned}$$

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

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 Φ, Θ — решение, то стохастические Φ', Θ' — тоже решения

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

Регуляризация — стандартный приём доопределения решения с помощью дополнительных критериев.

ARTM — Аддитивная Регуляризация Тематических Моделей

Максимизация log-правдоподобия с регуляризатором R:

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

EM-алгоритм: метод простой итерации для решения системы уравнений

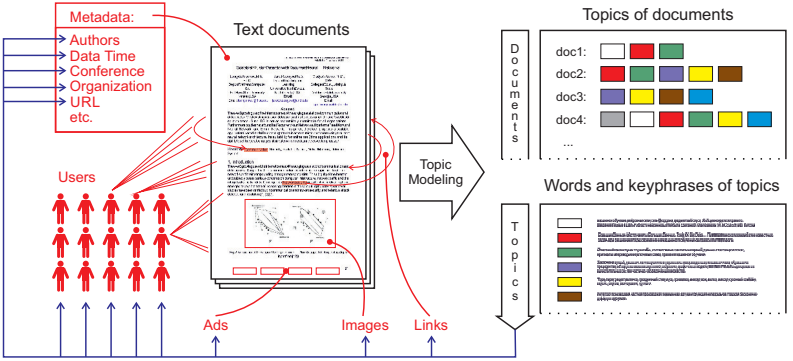
$$\text{E-шаг:} \quad p_{tdw} = \underset{t \in T}{\text{norm}}\left(\phi_{wt}\theta_{td}\right)$$

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

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

Задачи мультимодального тематического моделирования

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



Мультимодальная ARTM

Максимизация log-правдоподобий модальностей со словарями токенов W^m, $m \in M$:

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

EM-алгоритм: метод простой итерации для решения системы уравнений

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

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

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

Пример. Модальность n-грамм улучшает качество тем

Коллекция 1000 статей конференций ММРО, ИОИ на русском языке

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

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

BigARTM: библиотека тематического моделирования

Ключевые возможности:

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

Сообщество:

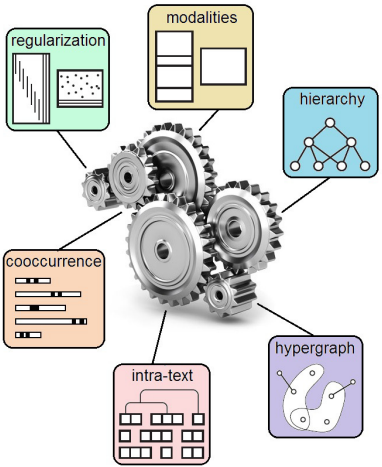
- Открытый код https://github.com/bigartm
  (discussion group, issue tracker, pull requests)
- Документация http://bigartm.org

Лицензия и среда разработки:

- Свободная коммерческая лицензия (BSD 3-Clause)
- Кросс-платформенность: Windows, Linux, MacOS (32/64 bit)
- Интерфейсы API: command-line, C++, and Python

Ключевые механизмы BigARTM

1. PLSA, LDA
2. регуляризация
3. модальности
4. иерархия тем
5. пост-обработка E-шага
6. совстречаемость термов
7. гиперграфы транзакций



Качество и скорость: BigARTM vs Gensim и Vowpal Wabbit

3.7M статей Википедии, 100K слов

		$T = 50$		$T = 200$	
	проц.	минут	перплексия	минут	перплексия
BigARTM	1	42	5117	83	3347
BigARTM async	1	25	5131	53	3362
VowpalWabbit	1	50	5413	154	3960
Gensim	1	142	4945	637	3241
BigARTM	4	12	5216	26	3520
BigARTM async	4	7	5353	16	3634
Gensim	4	88	5311	315	3583
BigARTM	8	8	5648	15	3929
BigARTM async	8	5	6220	10	4309
Gensim	8	88	6344	288	4263

D.Kochedykov, *M.Apishev*, *L.Golitsyn*, *K.Vorontsov*. Fast and Modular Regularized Topic Modelling. FRUCT ISMW, 2017.

Транзакционные данные

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

Примеры:

- **Данные социальной сети:**
  (d, u, w) — пользователь u записал слово w в блоге d
- **Данные сети интернет-рекламы:**
  (u, d, b) — пользователь u кликнул баннер b на странице d
- **Данные рекомендательной системы:**
  (u, f, s) — пользователь u оценил фильм f в ситуации s
- **Данные финансовых организаций:**
  (b, s, g) — покупатель u купил у продавца s товар g

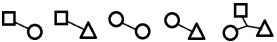
Задача: по выборке рёбер гиперграфа выявить латентные темы его вершин.

Тематическая модель гиперграфа: определения и обозначения

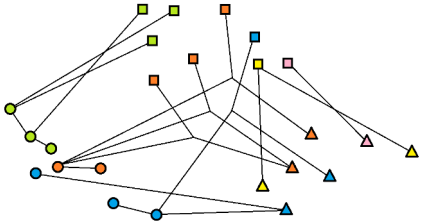
$\Gamma = \langle V, E \rangle$ — ориентированный гиперграф.
$V = V^1 \sqcup \cdots \sqcup V^M$ — разбиение вершин по модальностям

M — множество модальностей:

K — множество типов рёбер:



T — множество тем:



X^k — наблюдаемая выборка транзакций — рёбер типа k
ребро (d, x) состоит из вершины-контейнера $d \in V$ и множества вершин $x \subset V$,
n_{dx} — число вхождений ребра (d, x) в выборку X^k
$p(d, x)$ — неизвестное распределение на рёбрах типа k

Тематическая модель гиперграфа

Вероятностная тематическая модель рёбер типа k:

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

$\theta_{td} = p(t|d)$ — тематика контейнера не зависит от типа ребра k
$\phi_{vt} = p(v|t)$ — распределение термов модальности v в теме t

Задача максимизации log-правдоподобия:

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

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

где $\tau_k > 0$ — веса типов рёбер.

EM-алгоритм для гиперграфовой ARTM

Задача максимизации регуляризованного log-правдоподобия:

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

EM-алгоритм: метод простой итерации для решения системы уравнений

$$\begin{cases} \text{E-шаг:} & p_{tdx} = \underset{t\in T}{\text{norm}}\Big(\theta_{td}\prod_{v\in x}\phi_{vt}\Big) \\ \text{M-шаг:} & \phi_{vt} = \underset{v\in V^m}{\text{norm}}\Big(\sum_{k\in K}\tau_k \sum_{(d,x)} \big[v\in x\big] n_{dx}p_{tdx} + \phi_{vt}\frac{\partial R}{\partial \phi_{vt}}\Big) \\ & \theta_{td} = \underset{t\in T}{\text{norm}}\Big(\sum_{k\in K}\tau_k \sum_{(d,x)} n_{dx}p_{tdx} + \theta_{td}\frac{\partial R}{\partial \theta_{td}}\Big) \end{cases}$$

Направления текущих исследований в тематическом моделировании

- «внимание на лингвистику»: модели, учитывающие связность текста
- решение проблемы несбалансированности тем
- агрегирование гетерогенных коллекций
- автоматическая оптимизация коэффициентов регуляризации
- построение иерархических тематических векторов по Википедии
- графическая визуализация результатов тематического поиска
- детекция новых тем, в том числе в иерархических моделях
- автоматическое именование и суммаризация тем
- тематическая суммаризация произвольного набора
- тематическая сегментация документов

Литература


К.В.Воронцов. Обзор вероятностных тематических моделей. 2018. – NEW!
http://www.MachineLearning.ru/wiki/images/d/d5/Voron17survey-artm.pdf


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


K.Vorontsov, A.Potapenko. Additive regularization of topic models. Machine Learning, 2015.


O.Frei, M.Apishev. Parallel non-blocking deterministic algorithm for online topic modeling. AIST 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.


V.Alekseev, V.Bulatov, K.Vorontsov. Intra-Text Coherence as a Measure of Topic Models Interpretability. Dialogue, 2018.


A.Belyy, M.Seleznova, A.Sholokhov, K.Vorontsov. Quality Evaluation and Improvement for Hierarchical Topic Modeling. Dialogue, 2018.


N.Skachkov, K.Vorontsov. Improving topic models with segmental structure of texts. Dialogue, 2018.