

Тематический информационный поиск

Воронцов Константин Вячеславович
(Лаборатория машинного интеллекта МФТИ)

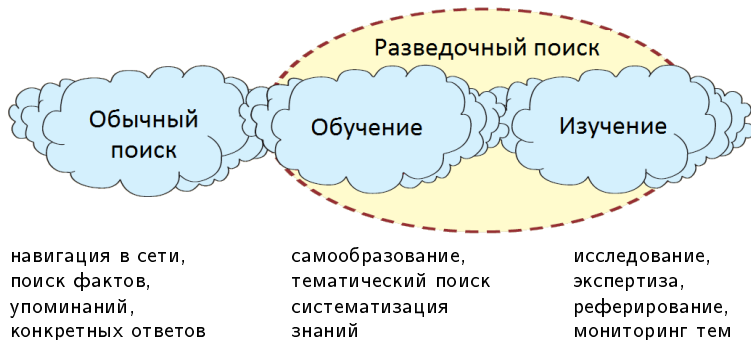
ФИЗТЕХ.РОСТ

- Открытая лекция от AI Business Course •
МФТИ • 18 октября 2019

- 1 От поиска информации к «мастерской знаний»**
 - Разведочный информационный поиск
 - Мастерская знаний и тематический поиск
 - Тематическое моделирование
- 2 Вероятностное тематическое моделирование**
 - Постановка задачи и классические методы
 - Теория аддитивной регуляризации
 - Библиотека тематического моделирования BigARTM
- 3 Примеры тематического анализа текстовых коллекций**
 - Этно-релевантные темы в социальных медиа
 - Интерпретируемость тем
 - Качество тематического поиска в экспериментах

Концепция разведочного поиска (exploratory search)

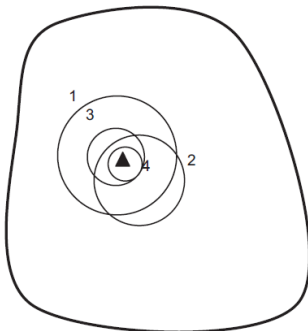
- пользователь может не знать ключевых терминов
- запросом может быть текст произвольной длины
- информационная потребность — систематизация знаний



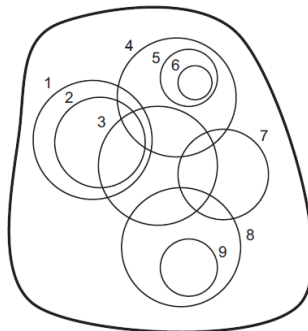
Gary Marchionini. Exploratory Search: from finding to understanding. 2006.

От итераций «query-browse-refine» к разведочному поиску

Iterative Search



Exploratory Search



▲ Search target



Information space



Result sets (larger = more results, intersection = overlap, # = iteration)

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

Обзор определений и моделей разведочного поиска

Основные свойства разведочного поиска:

- 1 An evolving search process**
разведочный поиск – это многошаговый процесс
каждый шаг – переформулировка или дополнение запроса
- 2 An anomalous state of knowledge**
в начале поиска у пользователя есть лишь мотивации,
но нет знаний и нет определённого плана, как их получать
- 3 Multiple targets / goals of search**
нет конкретной, точно определённой цели поиска
есть лишь общий интерес и эволюционирующие подцели

Обзор определений и моделей разведочного поиска

Свойства неопределённости процесса разведочного поиска

- 1 **Multiple possible answers**
возможных правильных ответов может быть много
- 2 **Not an expected exact answer**
не существует единственного правильного ответа
- 3 **A serendipitous attitude**
любой шаг может давать неожиданные новые знания
- 4 **An evolving information need**
на любом шаге цели и стратегии поиска могут измениться
- 5 **Uncertainty is fluctuating**
в процессе поиска неопределённость уменьшается,
но изменение цели может снова её увеличить

E.Palagi et al. A Survey of Definitions and Models of Exploratory Search. 2017.

Обзор определений и моделей разведочного поиска

Свойства разветвлённости процесса разведочного поиска

9 Multifaceted search

при поиске используются различные фильтры (фасеты)
примеры: по авторам, тематике, свежести, сложности

10 Several one-off pinpoint searches

многократные точечные одноразовые ответвления поиска
примеры: найти определение понятия, первоисточник

11 An open-ended search activity which can occur over time

процесс поиска никогда не заканчивается
пользователь может вернуться после долгого перерыва

Концепция «мастерской знаний»

Огромное и все возрастающее богатство знаний разбросано сегодня по всему миру. Этих знаний, вероятно, было бы достаточно для решения всего громадного количества трудностей наших дней, но они рассеяны и неорганизованы. Нам необходима очистка мышления в *своеобразной мастерской*, где можно **получать, сортировать, суммировать, усваивать, разъяснять и сравнивать** знания и идеи.
— Герберт Уэллс, 1940

An immense and ever-increasing wealth of knowledge is scattered about the world today; knowledge that would probably suffice to solve all the mighty difficulties of our age, but it is dispersed and unorganized. We need a sort of mental clearing house for the mind: a depot where knowledge and ideas are **received, sorted, summarized, digested, clarified and compared**
— Herbert Wells, 1940

Концепция тематического разведочного поиска

Поисковый запрос:

- документ любой длины или подборка документов

Результаты поиска:

- к каким темам относится мой запрос
- что ещё известно по этим темам
- в каком порядке читать, чтобы лучше разобраться в теме
- какие ещё темы часто встречаются вместе с данным
- какие фрагменты наиболее релевантны этим темам

Сценарий поиска:

- имея текст или подборку,
- получить рекомендации по расширению подборки,
- тематизировать, суммаризировать, визуализировать, . . .

Что такое «тема» в коллекции текстовых документов?

- *тема* — специальная терминология предметной области
- *тема* — набор часто совместно встречающихся терминов
- *тема* — семантически однородный кластер текстов

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

Имея коллекцию текстовых документов, хотим узнать:

- из каких тем состоит коллекция;
- из каких тем состоит каждый документ,
 $p(t|d)$ — вероятность темы t в документе d ;
- из каких слов или терминов состоит каждая тема,
 $p(w|t)$ — вероятность термина w в теме t .

Пример. Мультиязычная модель Википедии

216 175 русско-английских пар статей. Языки — модальности.
Первые 10 слов и их частоты $p(w|t)$ в %:

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

Ассессор оценил 396 тем из 400 как хорошо интерпретируемые.

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

Пример. Мультиязычная модель Википедии

216 175 русско-английских пар статей. Языки — модальности.
Первые 10 слов и их частоты $p(w|t)$ в %:

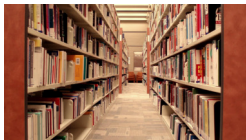
Тема №88				Тема №251			
opera	7.36	опера	7.82	windows	8.00	windows	6.05
conductor	1.69	оперный	3.13	microsoft	4.03	microsoft	3.76
orchestra	1.14	дирижер	2.82	server	2.93	версия	1.86
wagner	0.97	певец	1.65	software	1.38	приложение	1.86
soprano	0.78	певица	1.51	user	1.03	сервер	1.63
performance	0.78	театр	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

Ассессор оценил 396 тем из 400 как хорошо интерпретируемые.

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

Некоторые приложения тематического моделирования

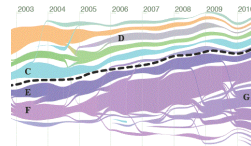
разведочный поиск в
электронных библиотеках



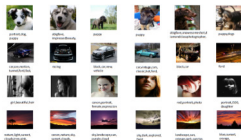
поиск тематического
контента в соцсетях



детектирование и трекинг
новостных сюжетов



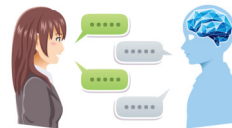
мультимодальный поиск
текстов и изображений



анализ банковских
транзакционных данных



управление диалогом в
разговорном интеллекте



Задача тематического моделирования

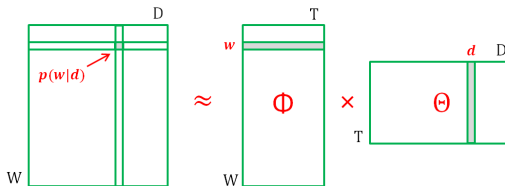
Дано: коллекция текстовых документов, $p(w|d) = \frac{n_{dw}}{n_d}$

Вероятностная тематическая модель:

$$p(w|d) = \sum_{t \in T} p(w|t)p(t|d) = \sum_{t \in T} \phi_{wt}\theta_{td}$$

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

Это задача стохастического матричного разложения:



Hofmann T. Probabilistic Latent Semantic Indexing. ACM SIGIR, 1999.

Blei D., Ng A., Jordan M. Latent Dirichlet Allocation. JMLR, 2003.

Принцип максимума правдоподобия

Правдоподобие — плотность распределения выборки $(d_i, w_i)_{i=1}^n$:

$$\prod_{i=1}^n p(d_i, w_i) = \prod_{d \in D} \prod_{w \in d} p(d, w)^{n_{dw}}$$

Максимизация логарифма правдоподобия

$$\sum_{d \in D} \sum_{w \in d} n_{dw} \ln p(w|d)p(d) \rightarrow \max_{\Phi, \Theta}$$

эквивалентна максимизации функционала

$$\mathcal{L}(\Phi, \Theta) = \sum_{d \in D} \sum_{w \in d} n_{dw} \ln \sum_{t \in T} \phi_{wt} \theta_{td} \rightarrow \max_{\Phi, \Theta}$$

при ограничениях неотрицательности и нормировки

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

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

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

Регуляризация — стандартный приём доопределения решения
с помощью дополнительных критериев.

ARTM: аддитивная регуляризация тематических моделей

Максимизация \log правдоподобия с регуляризатором R :

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

EM-алгоритм: метод простой итерации для системы уравнений

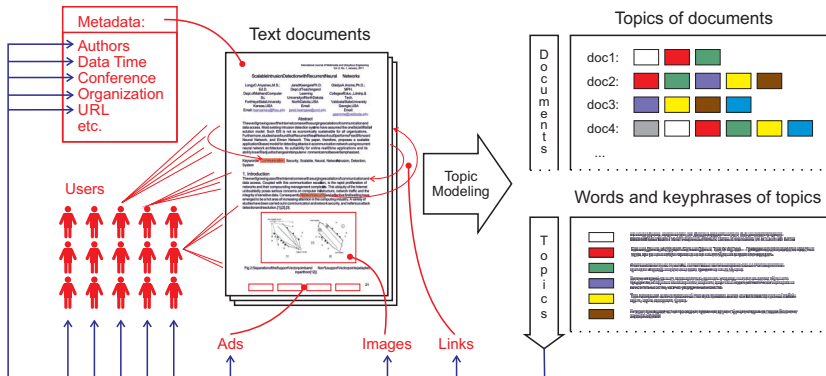
$$\begin{cases} \text{E-шаг:} & p_{tdw} \equiv p(t|d, w) = \operatorname{norm}_{t \in T}(\phi_{wt} \theta_{td}) \\ \text{M-шаг:} & \begin{cases} \phi_{wt} = \operatorname{norm}_{w \in W}(\sum_d n_{dw} p_{tdw} + \phi_{wt} \frac{\partial R}{\partial \phi_{wt}}) \\ \theta_{td} = \operatorname{norm}_{t \in T}(\sum_w n_{dw} p_{tdw} + \theta_{td} \frac{\partial R}{\partial \theta_{td}}) \end{cases} \end{cases}$$

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

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: $p(\text{автор}|t)$, $p(\text{время}|t)$, $p(\text{ссылка}|t)$, $p(\text{баннер}|t)$, $p(\text{элемент_изображения}|t)$, $p(\text{пользователь}|t)$, ...



Мультимодальная ARTM

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

Максимизация суммы \log правдоподобий с регуляризацией:

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

EM-алгоритм: метод простой итерации для системы уравнений

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

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

BigARTM: библиотека тематического моделирования

Ключевые возможности:

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

Сообщество:

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

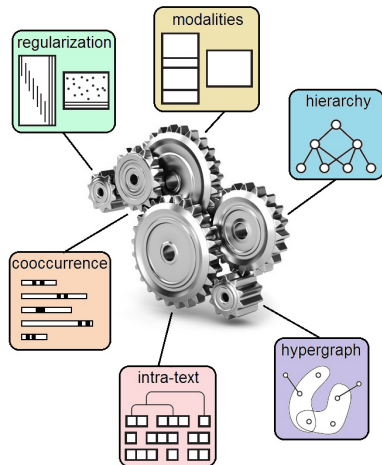


Лицензия и среда разработки:

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

Шесть ключевых механизмов BigARTM

- 1 регуляризация
- 2 модальности
- 3 иерархия тем
- 4 встречаемость термов
- 5 внутритекстовые связи
- 6 гиперграфовые данные



Качество и скорость: BigARTM vs Gensim и Vowpal Wabbit

3.7М статей Википедии, 100К слов:

время min (перплексия)

проц.	$ T $	Gensim	Vowpal Wabbit	BigARTM	BigARTM асинхрон
1	50	142m (4945)	50m (5413)	42m (5117)	25m (5131)
1	100	287m (3969)	91m (4592)	52m (4093)	32m (4133)
1	200	637m (3241)	154m (3960)	83m (3347)	53m (3362)
2	50	89m (5056)		22m (5092)	13m (5160)
2	100	143m (4012)		29m (4107)	19m (4144)
2	200	325m (3297)		47m (3347)	28m (3380)
4	50	88m (5311)		12m (5216)	7m (5353)
4	100	104m (4338)		16m (4233)	10m (4357)
4	200	315m (3583)		26m (3520)	16m (3634)
8	50	88m (6344)		8m (5648)	5m (6220)
8	100	107m (5380)		10m (4660)	6m (5119)
8	200	288m (4263)		15m (3929)	10m (4309)

D.Kochedykov, M.Apishev, L.Golitsyn, K.Vorontsov.

Fast and Modular Regularized Topic Modelling. FRUCT ISMW, 2017.

BigARTM упрощает разработку тематических моделей

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

Этапы моделирования

Bayesian TM

ARTM

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

-- нестандартизируемые этапы, уникальная разработка для каждой задачи

-- стандартизируемые этапы

Байесовское обучение — доминирующий подход в ТМ

Основа подхода — байесовский вывод:

$$\text{Posterior}(\Phi, \Theta | \text{data}) \propto \text{Prior}(\Phi, \Theta) P(\text{data} | \Phi, \Theta)$$

В модели LDA Prior и Posterior — распределения Дирихле.

Проблемы:

- Нам нужны лишь значения Φ, Θ , а не их распределения
- Prior Дирихле имеет слабые лингвистические обоснования
- Задача сильно усложняется для несопряжённых Prior
- Байесовский вывод уникален для каждой модели
- Невозможно модульное комбинирование моделей

Blei D., Ng A., Jordan M. Latent Dirichlet Allocation. JMLR, 2003.

Поиск этно-релевантных тем в социальных сетях

Основные задачи проекта:

- Разведочный поиск этнических тем в социальных медиа (сколько различных тем, и что это за темы)
- Мониторинг этих тем во времени и по регионам
- Сентимент-анализ и оценивание конфликтности

Вспомогательные задачи:

- Фильтрация (обогащение) потока данных
- Обеспечение полноты поиска этнических тем
- Выявление тематических сообществ
- Выделение событийных и региональных тем
- Решение проблемы коротких сообщений

Примеры этнонимов

османский

восточноевропейский

эвенк

швейцарская

аланский

саамский

латыш

литовец

цыганка

ханты-мансийский

карачаевский

кубинка

гагаузский

русич

сингапурец

перуанский

словенский

вепсский

ниггер

адыги

сомалиец

абхаз

темнокожий

нигериец

лягушатник

камбоджиец

Примеры этнических тем

(русские): русский, князь, россия, татарин, великий, царить, царь, иван, император, империя, грозить, государь, век, московская, екатерина, москва,

(русские): акция, организация, митинг, движение, активный, мероприятие, совет, русский, участник, москва, оппозиция, россия, пикет, протест, проведение, националист, поддержка, общественный, проводить, участие,

(славяне, византийцы): славянский, святослав, жрец, древние, письменность, рюрик, летопись, византия, мефодий, хазарский, русский, азбука,

(сирийцы): сирийский, асад, боевик, район, террорист, уничтожать, группировка, дамаск, оружие, алесию, оппозиция, операция, селение, сша, нусра, турция,

(турки): турция, турецкий, курдский, эрдоган, стамбул, страна, кавказ, горин, полиция, премьер-министр, регион, курдистан, ататюрк, партия,

(иранцы): иран, иранский, сша, россия, ядерный, президент, тегеран, сирия, оон, израиль, переговоры, обама, санкция, исламский,

(палестинцы): террорист, израиль, терять, палестинский, палестинец, террористический, палестина, взрыв, территория, страна, государство, безопасность, арабский, организация, иерусалим, военный, полиция, газ,

(ливанцы): ливанский, боевик, район, ливан, армия, террорист, али, военный, хизбалла, раненый, уничтожать, сирия, подразделение, квартал, армейский,

(ливийцы): ливан, демократия, страна, ливийский, каддафи, государство, алжир, война, правительство, сша, арабский, али, муаммар, сирия,

Примеры этнических тем

(евреи): израиль, израильский, страна, война, нетаньяху, тель-авив, время, сша, сирия, египет, случай, самолет, еврейский, военный, ближний,

(американцы): американский, американка, война, россия, военный, страна, вашингтон, америка, армия, конгресс, сирия, союзный, российский, обама, войска, русский, оружие, операция,

(немцы): армия, война, войска, советский, военный, дивизия, немец, фронт, немецкий, генерал, борт, операция, оборона, русский, бог, победа,

(немцы): германий, немец, германский, ссср, немецкий, война, старое, советский, россия, береза, русский, правительство, территория, полный, документ, вопрос, сорт, договор, отношение, франция,

(евреи, немцы): еврей, еврейский, холодный, германий, антисемитизм, гетра, немец, синагога, сша, израиль, малиновского, комиссия, нацбол, документ, война, еврейка, миллион, украина,

(украинцы, немцы): украинский, унс, оун, немец, немецкий, ковальков, хохол, волынский, бандера, организация, россиянин, советский, русский, польский, армия, шухевича, ровенский,

(таджики, узбеки): мигрант, страна, россия, миграция, азия, нелегальный, миграционный, таджикистан, гастарбайтер, гражданка, трудовой, рабочий, фмс, коренево, среднее, узбекистан, таджик, проблема, русский, население,

(канадцы): команда, игра, игрок, канадский, сезон, хоккей, сборная, играть, болельщик, победа, кубок, счет, забирать, хоккейный, выигрывать, хоккеист, чемпионат, шайба,

Примеры этнических тем

(японцы): японский, япония, корея, китайский, жилища, авария, фукусиму, цунами, общаться, океан, станция, хатико, район, правительство, атомный,

(норвежцы): дитя, ребенок, родиться, детский, семья, воспитанный, право, возраст, отец, воспитание, норвежский, родительский, родить, мальчик, взрослый, опека, сын,

(венесуэльцы): куба, кастро, венесуэла, чавес, президент, уго, мадура, боливия, фидель, глава, латинский, венесуэльский, лидер, боливарианской, президентский, альенде, гевару,

(китайцы): китайский, россия, производство, китаи, продукция, страна, предприятие, компания, технология, военный, регион, производить, производственный, промышленность, российский, экономический, кнр,

(азербайджанцы): русский, азербайджан, азербайджанец, россия, азербайджанский, таксист, диаспора, анапа, народ, москва, страна, армянин, слово, рынок,

(грузины): грузинский, спецназ, военный, август, баташева, российский, спецназовец, миротворец, операция, румын, бригада, миротворческий, абхазия, группа, войска, русский, цхинвале,

(осетины): конституция, осетия, аминат, русский, осетинский, южный, северный, россия, война, республика, вопрос, алахай, российский, население, конфликт,

(цыгане): наркотик, цыган, цыганка, хороший, место, страна, деньга, время, работать, жизнь, жить, рука, дом, цыганский, наркоманка,

Результаты: модель ARTM находит намного больше этно-тем

Число этно-релевантных тем, найденных моделью:

модель	этно-тем	фон.тем	++	+-	-+	всего
PLSA	300		9	11	18	38
PLSA	400		12	15	17	44
ARTM-1	200	100	18	33	20	71
ARTM-1	250	150	21	27	20	68
ARTM-2	200	100	28	23	23	74
ARTM-2	250	150	38	42	30	104

Регуляризаторы ARTM-1:

этно темы: разреживание, декоррелирование, сглаживание этнонимов

фоновые темы: сглаживание, разреживание этнонимов

Регуляризаторы ARTM-2:

ARTM-1 + **модальность этнонимов**

Мультиязычная модель Википедии

216 175 русско-английских пар статей Вики.

Первые 10 слов и их вероятности $p(w|t)$ 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

Дударенко М. А. Регуляризация многоязычных тематических моделей. Вычислительные методы и программирование. 2015. Т. 16. С. 26–36.

Мультиязычная модель Википедии

216 175 русско-английских пар статей Вики.

Первые 10 слов и их вероятности $p(w|t)$ 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 ассессор оценил 396 тем из $|T| = 400$ как хорошо интерпретируемые.

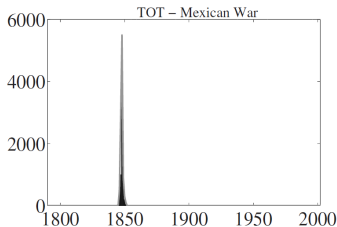
Биграммы радикально улучшают интерпретируемость тем

Коллекция 1000 статей конференций ММРО, ИОИ на русском

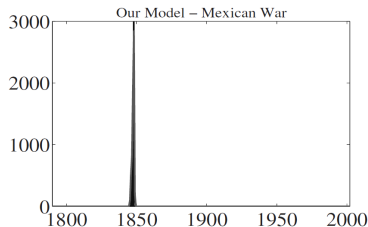
распознавание образов в биоинформатике		теория вычислительной сложности	
unigrams	bigrams	unigrams	bigrams
объект	задача распознавания	задача	разделять множества
задача	множество мотивов	множество	конечное множество
множество	система масок	подмножество	условие задачи
мотив	вторичная структура	условие	задача о покрытии
разрешимость	структура белка	класс	покрытие множества
выборка	распознавание вторичной	решение	сильный смысл
маска	состояние объекта	конечный	разделяющий комитет
распознавание	обучающая выборка	число	минимальный аффинный
информативность	оценка информативности	аффинный	аффинный комитет
состояние	множество объектов	случай	аффинный разделяющий
закономерность	разрешимость задачи	покрытие	общее положение
система	критерий разрешимости	общий	множество точек
структура	информативность мотива	пространство	случай задачи
значение	первичная структура	схема	общий случай
регулярность	тупиковое множество	комитет	задача MASC

Совмещение динамической и n -граммной модели

По коллекции выступлений президентов США



1. mexico	8. territory
2. texas	9. army
3. war	10. peace
4. mexican	11. act
5. united	12. policy
6. country	13. foreign
7. government	14. citizens

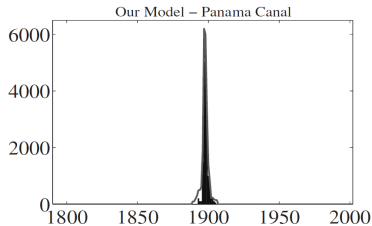
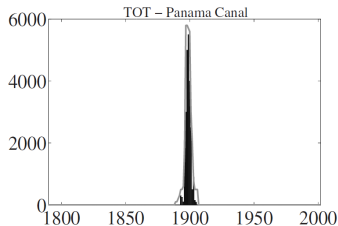


1. east bank	8. military
2. american coins	9. general herrera
3. mexican flag	10. foreign coin
4. separate independent	11. military usurper
5. american commonwealth	12. mexican treasury
6. mexican population	13. invaded texas
7. texan troops	14. veteran troops

Shoaib Jameel, Wai Lam. An N -Gram Topic Model for Time-Stamped Documents. ECIR 2013.

Совмещение динамической и n -граммной модели

По коллекции выступлений президентов США



1. government	8. spanish
2. cuba	9. island
3. islands	10. act
4. international	11. commission
5. powers	12. officers
6. gold	13. spain
7. action	14. rico

1. panama canal	8. united states senate
2. isthmian canal	9. french canal company
3. isthmus panama	10. caribbean sea
4. republic panama	11. panama canal bonds
5. united states government	12. panama
6. united states	13. american control
7. state panama	14. canal

Shoaib Jameel, Wai Lam. An N-Gram Topic Model for Time-Stamped Documents. ECIR 2013.

Две коллекции новостей про технологии

Habr.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

Предобработка текстов

- отброшены 5% наиболее частотных слов (общая лексика)
- удаление пунктуации
- нижний регистр, ё→е
- лемматизация r morphology2

Методика оценивания качества разведочного поиска

Поисковый запрос

набор ключевых слов или фрагментов текста, около одной страницы A4

Поисковая выдача

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

Два задания асессорам

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

Поиск MapReduce

Поиск MapReduce – программа поиска (поисковик) написана распределенно: выделены для больших объемов данных и разнородных запросов, представляющих собой набор Java-классов и исполняемых утилит для создания и обработки данных на параллельно обрабатываемых серверах.

Основные компоненты Поиска MapReduce можно сформулировать как:

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

Поиск – популярная программная платформа (**поиск**, **поисковик**) построена распределенными приложениями для массово-параллельной обработки (**раздел**, **раздел**, **раздел**, **МР**) данных.

Поиск включает в себе следующие компоненты:

1. **MRFS** – распределенная файловая система;
2. **Поиск MapReduce** – программная платформа (**поиск**, **поисковик**) написана распределенно: выделены для больших объемов данных и разнородных запросов.

Ключевые, **облачные** и архитектура **Поиска MapReduce** и структура **MRFS** стали примером того, как можно работать с данными, в том числе и с помощью точки зрения **MRFS**, и в конечном итоге, определило ограничение платформ **Поиск** и в целом **MRFS** пользователи можно отметить:

Ограничение масштабируемости кластера **Поиск** – не масштабируемый утилит, «не масштабируемый задания».

Сильная зависимость **Поиска** распределенно вычислений и клиентских выделены, реализованы распределенно алгоритмы. Как следствие:

Отсутствие поддержки алгоритмической программы выделены распределенно вычислений в **Поиск** v1.0 поддерживается только только выделены **раздел**.

Многие выделены, точки зрения и как следствие, масштабируемость выделены в средстве с выделены требования в масштабируемости.

Проблема **выделены** совместности требования по **выделены** выделены выделены все выделены утилит кластера при **выделены** платформ **Поиск** (установка **выделены** или **выделены** **выделены**).

Пример запроса для разведочного поиска

Пример: фрагмент запроса «Система IBM Watson»

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

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

... ..

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

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

Тематика запросов разведочного поиска

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

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

Оценивание качества поиска

Precision — доля релевантных среди найденных

Recall — доля найденных среди релевантных

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

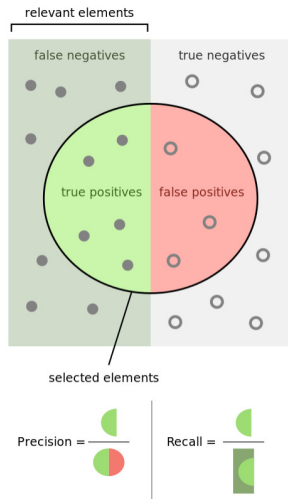
$$R = \frac{TP}{TP + FN} \text{ — полнота, (recall)}$$

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

TP (true positive) — найденные релевантные

FP (false positive) — найденные нерелевантные

FN (false negative) — не найденные релевантные



Какие модели поиска сравнивались

- **assessors**: результаты поиска, выполненного ассессорами
- **TF-IDF, BM25**: сравнение документов по частотам слов
- **word2vec**: нетематические векторные представления слов
- **PLSA**: Probabilistic Latent Semantic Analysis (1999)
- **LDA**: Latent Dirichlet Allocation (2003)
- **ARTM**: тематическая модель с тремя регуляризаторами
- **hARTM**: двухуровневая иерархическая модель ARTM

Задачи регуляризаторов в ARTM и hARTM:

- сделать темы как можно более различными
- сделать векторы $p(t|d)$ как можно более разреженными
- не допустить вырождения распределений $p(w|t)$

Стратегия регуляризации

Последовательное применение трёх регуляризаторов

- 1 декоррелирование тем:

$$R(\Phi) = -\tau \sum_{s,t \in T} \sum_{w \in W} \phi_{wt} \phi_{ws}$$

- 2 разреживание распределений $p(t|d)$:

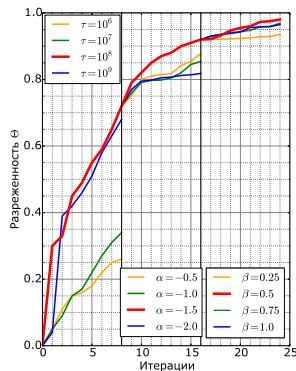
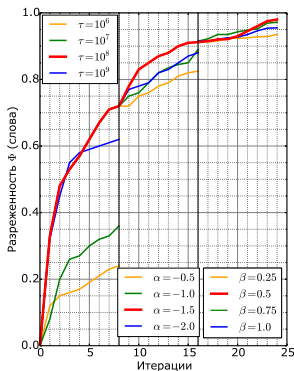
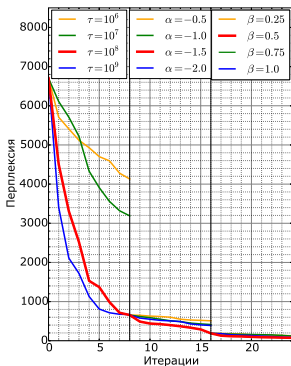
$$R(\Theta) = -\alpha \sum_{d,t} \ln \theta_{td}$$

- 3 сглаживание распределений $p(w|t)$:

$$R(\Phi) = \beta \sum_{t,w} \ln \phi_{wt}$$

Последовательный подбор коэффициентов регуляризации

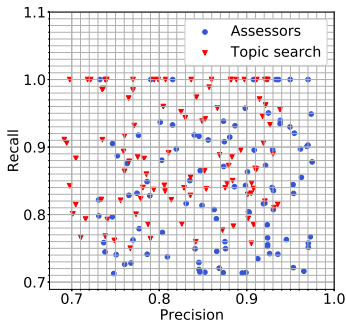
- декоррелирование распределений термов в темах (τ),
- разреживание распределений тем в документах (α),
- сглаживание распределений термов в темах (β).



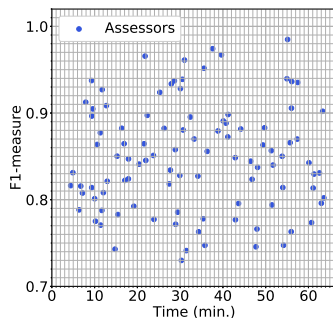
Результаты измерения точности и полноты по запросам

100 запросов, 3 ассессора на запрос

точность и полнота поиска



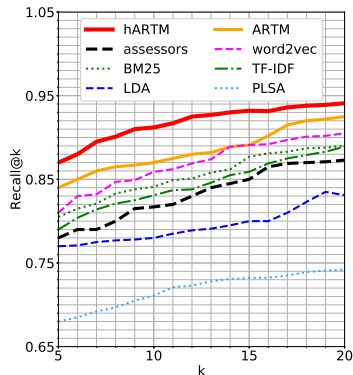
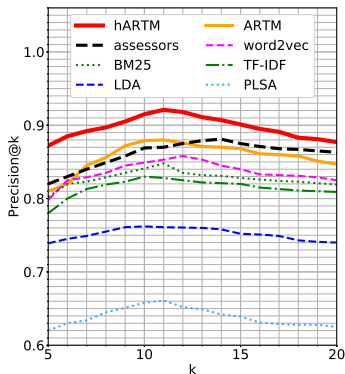
время и F_1 -мера (ассессоры)



- среднее время обработки запроса ассессором — 30 минут
- точность выше у ассессоров, полнота — у поисковика

Сравнение с ассессорами по качеству поиска

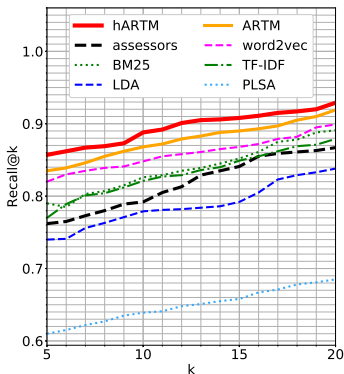
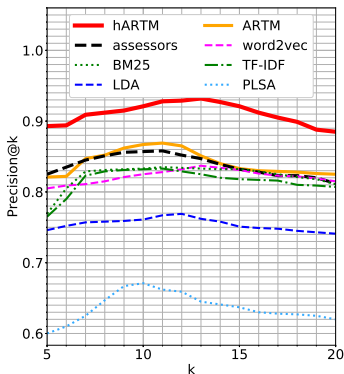
Точность и полнота по первым k позициям поисковой выдачи (коллекция Habrahabr.ru)



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

Сравнение с ассессорами по качеству поиска

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



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

Влияние числа тем на качество поиска

Все регуляризаторы и модальности, **плоская модель**

	Habrahabr						TechCrunch					
	асесс	100	150	200	250	400	асесс	350	400	450	475	500
Pr@5	0.821	0.662	0.721	0.810	0.761	0.693	0.822	0.653	0.725	0.752	0.819	0.777
Pr@10	0.869	0.761	0.812	0.879	0.825	0.673	0.851	0.663	0.732	0.762	0.867	0.811
Pr@15	0.875	0.733	0.795	0.868	0.791	0.651	0.835	0.682	0.743	0.787	0.833	0.793
Pr@20	0.863	0.724	0.795	0.847	0.792	0.642	0.813	0.650	0.743	0.773	0.825	0.793
R@5	0.780	0.732	0.807	0.840	0.821	0.721	0.762	0.731	0.762	0.793	0.835	0.817
R@10	0.817	0.771	0.843	0.870	0.851	0.751	0.792	0.763	0.793	0.812	0.868	0.855
R@15	0.850	0.824	0.895	0.891	0.871	0.773	0.835	0.782	0.807	0.855	0.890	0.882
R@20	0.873	0.857	0.905	0.925	0.892	0.771	0.867	0.792	0.823	0.862	0.919	0.903

- существует оптимальное число тем
- чем больше коллекция, тем больше оптимум числа тем

Влияние числа тем на качество поиска

Nabrahabr. Все регуляризаторы и модальности, **два уровня**

$ T_1 $	20		25			30					
$ T_2 $	150	200	250	275	300	400	450				
Pr@5	0.621	0.742	0.839	0.850	0.865	0.869	0.869	0.803	0.769	0.701	0.670
Pr@10	0.645	0.749	0.850	0.861	0.879	0.911	0.895	0.809	0.796	0.719	0.689
Pr@15	0.635	0.751	0.848	0.869	0.873	0.893	0.887	0.807	0.781	0.721	0.701
Pr@20	0.630	0.745	0.841	0.855	0.864	0.874	0.875	0.800	0.775	0.709	0.675
R@5	0.628	0.773	0.843	0.865	0.881	0.881	0.868	0.849	0.839	0.715	0.691
R@10	0.652	0.782	0.855	0.871	0.902	0.918	0.877	0.871	0.845	0.745	0.699
R@15	0.671	0.801	0.870	0.889	0.929	0.939	0.901	0.883	0.861	0.781	0.722
R@20	0.680	0.819	0.886	0.892	0.955	0.955	0.907	0.901	0.872	0.801	0.729

- существует оптимальное число тем на каждом уровне
- два уровня лучше, чем один
- увеличивается оптимальное число тем на нижнем уровне

Влияние числа тем на качество поиска

Nabrahabr. Все регуляризаторы и модальности, **три уровня**

$ T_1 $	20		25						30		
$ T_2 $	150	200	250		275		300		400	450	
$ T_3 $	750	800	1200	1300	1300	1400	1500	1500	1600	3000	3500
Pr@5	0.625	0.743	0.840	0.852	0.869	0.872	0.870	0.805	0.771	0.705	0.672
Pr@10	0.648	0.754	0.851	0.867	0.882	0.915	0.901	0.811	0.799	0.722	0.694
Pr@15	0.632	0.752	0.850	0.872	0.878	0.895	0.889	0.809	0.785	0.729	0.703
Pr@20	0.629	0.745	0.845	0.861	0.871	0.877	0.882	0.803	0.778	0.710	0.681
R@5	0.632	0.780	0.845	0.869	0.883	0.889	0.872	0.851	0.841	0.721	0.695
R@10	0.654	0.792	0.859	0.873	0.905	0.922	0.881	0.873	0.850	0.749	0.703
R@15	0.675	0.805	0.874	0.892	0.932	0.942	0.905	0.889	0.863	0.787	0.725
R@20	0.684	0.824	0.889	0.901	0.958	0.961	0.912	0.904	0.878	0.805	0.734

- существует оптимальное число тем на каждом уровне
- три уровня лучше, чем один или два
- увеличивается оптимальное число тем на нижнем уровне

Влияние числа тем на качество поиска

TechCrunch. Все регуляризаторы и модальности, **два уровня**

$ T_1 $	80		100			120					
$ T_2 $	300	350	500	550	600	700	750				
Pr@5	0.651	0.701	0.749	0.789	0.883	0.889	0.889	0.785	0.721	0.701	0.675
Pr@10	0.675	0.709	0.771	0.821	0.891	0.918	0.902	0.803	0.738	0.718	0.691
Pr@15	0.687	0.712	0.773	0.827	0.899	0.919	0.905	0.817	0.741	0.721	0.701
Pr@20	0.683	0.707	0.759	0.815	0.885	0.888	0.895	0.805	0.732	0.716	0.679
R@5	0.749	0.791	0.801	0.854	0.868	0.875	0.861	0.849	0.829	0.731	0.701
R@10	0.765	0.809	0.823	0.873	0.890	0.904	0.875	0.867	0.835	0.745	0.708
R@15	0.771	0.820	0.841	0.882	0.909	0.921	0.895	0.890	0.848	0.769	0.717
R@20	0.778	0.825	0.851	0.887	0.928	0.942	0.929	0.901	0.869	0.785	0.728

- существует оптимальное число тем на каждом уровне
- два уровня лучше, чем один
- увеличивается оптимальное число тем на нижнем уровне

Влияние числа тем на качество поиска

TechCrunch. Все регуляризаторы и модальности, **три уровня**

$ T_1 $	80		100						120		
$ T_2 $	300	350	500		550		600		700	750	
$ T_3 $	1500	1700	2500	2600	2600	2800	3000	3000	3200	4500	4700
Pr@5	0.655	0.707	0.751	0.792	0.887	0.893	0.890	0.789	0.722	0.703	0.678
Pr@10	0.678	0.712	0.773	0.823	0.895	0.922	0.905	0.805	0.741	0.722	0.692
Pr@15	0.692	0.715	0.775	0.831	0.902	0.921	0.907	0.821	0.743	0.725	0.703
Pr@20	0.687	0.709	0.761	0.819	0.889	0.885	0.898	0.809	0.736	0.719	0.683
R@5	0.751	0.795	0.802	0.856	0.871	0.877	0.863	0.852	0.831	0.738	0.705
R@10	0.767	0.812	0.825	0.875	0.892	0.908	0.879	0.871	0.842	0.751	0.711
R@15	0.772	0.824	0.841	0.887	0.912	0.927	0.901	0.893	0.854	0.772	0.721
R@20	0.783	0.830	0.854	0.892	0.931	0.949	0.935	0.905	0.871	0.790	0.732

- существует оптимальное число тем на каждом уровне
- три уровня лучше, чем один или два
- увеличивается оптимальное число тем на нижнем уровне

Влияние модальностей на качество поиска

Все регуляризаторы и модальности, 3 уровня, оптимальное $|T|$

Модальности: Words, Bigrams, Authors, Comments, Tags, Hubs, Categories

	Habrahabr						TechCrunch					
	асесс	W	Com	WB	WBTH	All	асесс	W	C	WB	WBC	All
Pr@5	0.821	0.621	0.558	0.673	0.871	0.872	0.822	0.718	0.569	0.795	0.891	0.893
Pr@10	0.869	0.645	0.567	0.712	0.911	0.915	0.851	0.729	0.592	0.807	0.919	0.922
Pr@15	0.875	0.631	0.532	0.693	0.894	0.895	0.835	0.737	0.603	0.803	0.920	0.921
Pr@20	0.863	0.628	0.531	0.688	0.877	0.877	0.813	0.729	0.594	0.792	0.883	0.885
R@5	0.780	0.725	0.645	0.797	0.888	0.889	0.762	0.754	0.659	0.775	0.874	0.877
R@10	0.817	0.748	0.652	0.812	0.921	0.922	0.792	0.778	0.671	0.808	0.908	0.908
R@15	0.850	0.782	0.679	0.842	0.941	0.942	0.835	0.783	0.679	0.825	0.927	0.927
R@20	0.873	0.789	0.672	0.852	0.960	0.961	0.867	0.785	0.711	0.837	0.949	0.949

- лучше использовать все модальности
- биграммы и категории выигрывают у ассессоров
- авторы и комментаторы наименее важны

Влияние регуляризаторов на качество поиска

Все регуляризаторы и модальности, 3 уровня, оптимальное |T|

Регуляризаторы: Decorrelation, Θ-sparsing, Φ-smoothing, Hierarchy

	Habrahabr					TechCrunch				
	нет	D	DΘ	DΘΦ	DΘΦH	нет	D	DΘ	DΘΦ	DΘΦH
Pr@5	0.628	0.772	0.771	0.865	0.872	0.652	0.777	0.779	0.879	0.893
Pr@10	0.653	0.781	0.812	0.883	0.915	0.679	0.788	0.819	0.895	0.922
Pr@15	0.642	0.785	0.792	0.891	0.895	0.669	0.791	0.798	0.901	0.921
Pr@20	0.643	0.771	0.783	0.875	0.877	0.673	0.775	0.792	0.892	0.885
R@5	0.692	0.820	0.805	0.875	0.889	0.673	0.825	0.812	0.869	0.877
R@10	0.714	0.831	0.834	0.905	0.922	0.685	0.856	0.845	0.881	0.908
R@15	0.725	0.847	0.867	0.921	0.942	0.712	0.877	0.869	0.912	0.927
R@20	0.735	0.873	0.891	0.943	0.961	0.723	0.892	0.895	0.934	0.949

- Лучше использовать все регуляризаторы
- Модели со слабой регуляризацией (PLSA, LDA) слабы

Влияние функции близости на качество поиска

Все регуляризаторы и модальности, 3 уровня, оптимальное $|T|$

Функции близости: Euclidean, Cosine, Manhattan, Hellinger, KL-div

	Habrahabr					TechCrunch				
	Eu	cos	Ma	He	KL	Eu	cos	Ma	He	KL
Pr@5	0.652	0.872	0.772	0.725	0.741	0.647	0.893	0.752	0.742	0.735
Pr@10	0.693	0.915	0.798	0.749	0.772	0.658	0.922	0.794	0.758	0.751
Pr@15	0.695	0.895	0.803	0.737	0.751	0.672	0.921	0.801	0.745	0.742
Pr@20	0.671	0.877	0.789	0.731	0.738	0.652	0.885	0.793	0.739	0.738
R@5	0.693	0.889	0.721	0.742	0.833	0.688	0.877	0.708	0.733	0.858
R@10	0.715	0.922	0.732	0.775	0.868	0.692	0.908	0.715	0.753	0.872
R@15	0.732	0.942	0.739	0.791	0.892	0.724	0.927	0.719	0.785	0.895
R@20	0.741	0.961	0.721	0.812	0.902	0.732	0.949	0.711	0.808	0.901

- косинусная функция близости уверенно лидирует

Выводы по результатам экспериментов

- Ассессорские данные относятся не к темам, а к коллекции; поэтому с их помощью можно оценивать новые модели
- Небольших ассессорских данных хватает для оценивания тематических моделей, которые обучаются *без учителя*
- Регуляризаторы, улучшающие интерпретируемость модели, повышают также и качество поиска
- Иерархия улучшает качество поиска (в основном точность) благодаря сужению области поиска
- Подбор траектории регуляризации и оптимизация коэффициентов регуляризации влияет на качество поиска
- При тщательной оптимизации тематический поиск превосходит как ассессоров, так и конкурирующие модели

Поисково-рекомендательная система arXiv.AITHEA.com











Пользователь собирает тематические подборки документов

The screenshot shows a web browser window displaying the arXiv.AITHEA.com website. The URL in the address bar is <https://arxiv.aithea.com/collections/Q29sbGVjdGVhbjotJFVTUEFkaHBH>. The website has a dark blue header with navigation tabs: FEEDS, SEARCH, and COLLECTIONS. The current page is titled "Topic Modeling for Opinion Mining" and is categorized under "PAPERS". There are two columns: "PAPERS" and "RECOMMENDED".

The first paper listed is dated 24 DEC 2017 and titled "Comparative Opinion Mining: A Review" by Kasturi Dewl Varathan, Anastasia Giachanou, and Fabio Crestani. The abstract states: "Opinion mining refers to the use of natural language processing, text analysis and computational linguistics to identify and extract subjective information in textual material. Opinion mining, also known as sentiment analysis, has received a lot of attention in recent times, as it provides a number of tools to analyse the public opinion on a number of different topics. Comparative opinion mining is a subfield of opinion mining that deals with identifying and extracting information that is expressed in a". It has 6 citations.

The second paper is dated 7 NOV 2015 and titled "The survey of sentiment and opinion mining for behavior analysis of social media" by Saqib Iqbal, Ali Zulqurnain, Yaqoob Wani, and Khalid Hussain. The abstract states: "Nowadays, internet has changed the world into a global village. Social Media has reduced the gaps among the individuals. Previously communication was a time consuming and expensive task between the people. Social Media has earned fame because it is a cheaper and faster communication provider. Besides, social media has allowed us to reduce the gaps of physical distance, it also generates and preserves huge amount of data. The data are very valuable and it presents association degree". It has 0 citations.

Разработка компании Digital Decisions (AITHEA)

-  *К.В.Воронцов*. Обзор вероятностных тематических моделей. 2017. – **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.