Тематическое моделирование для поиска и систематизации научно-технической информации

Воронцов Константин Вячеславович
ФИЦ ИУ РАН • МФТИ • МГУ •

«Информационные технологии в современной библиотеке»
13 сентября 2016

Аннотация

Технологии информационного поиска сделали научные знания доступнее. Их стало легче находить, но это не означает, что их стало легче понимать. Ответ на вопрос «где находится передний край науки по данной теме» по-прежнему требует времени, квалификации и личного общения с экспертами.

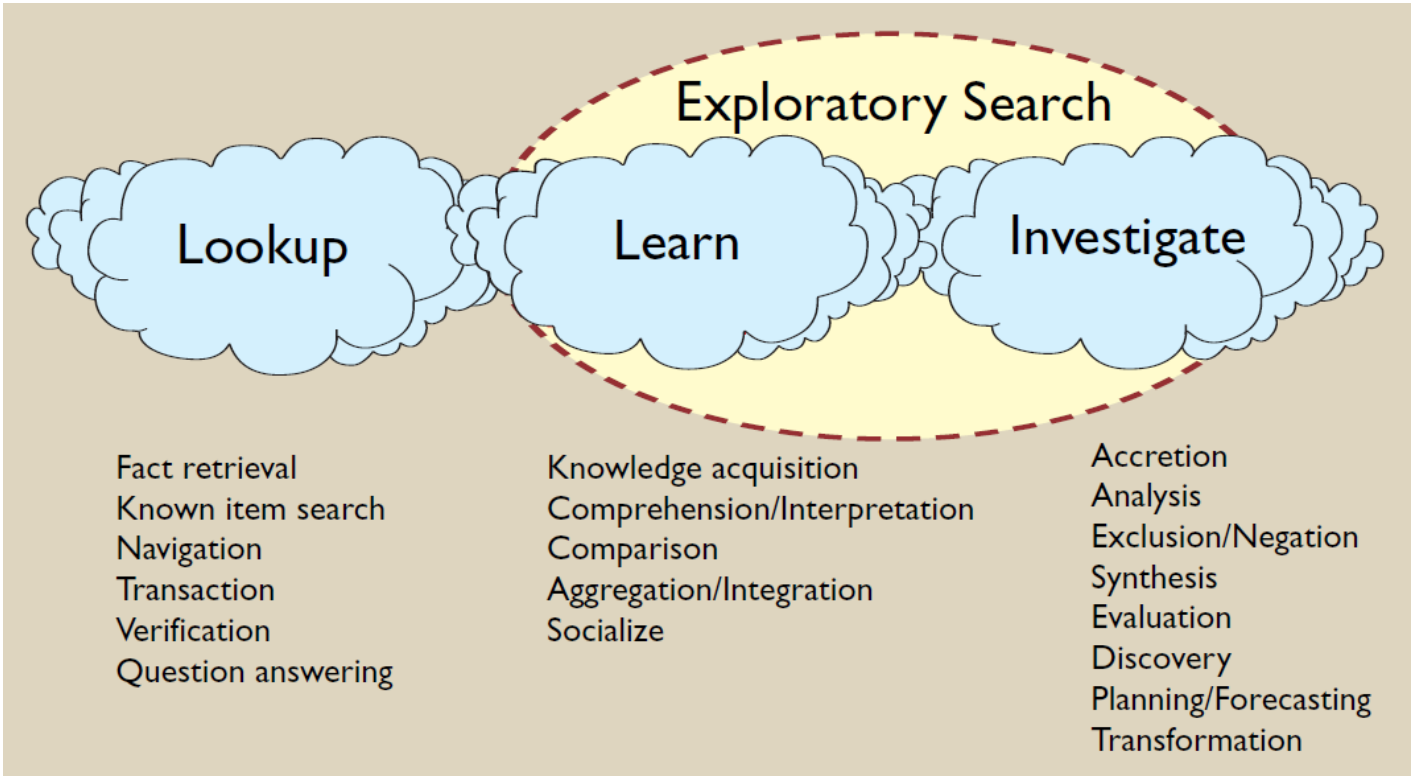
Разведочный поиск (exploratory search) — это новая парадигма в информационном поиске, нацеленная на дальнейшее устранение барьеров между Человеком и Знанием. Разведочный поиск призван объединить и автоматизировать процессы поиска, систематизации и усвоения знаний.

В докладе рассматриваются методы *вероятностного тематического моделирования* больших текстовых коллекций и их применение для тематического разведочного поиска.

Содержание

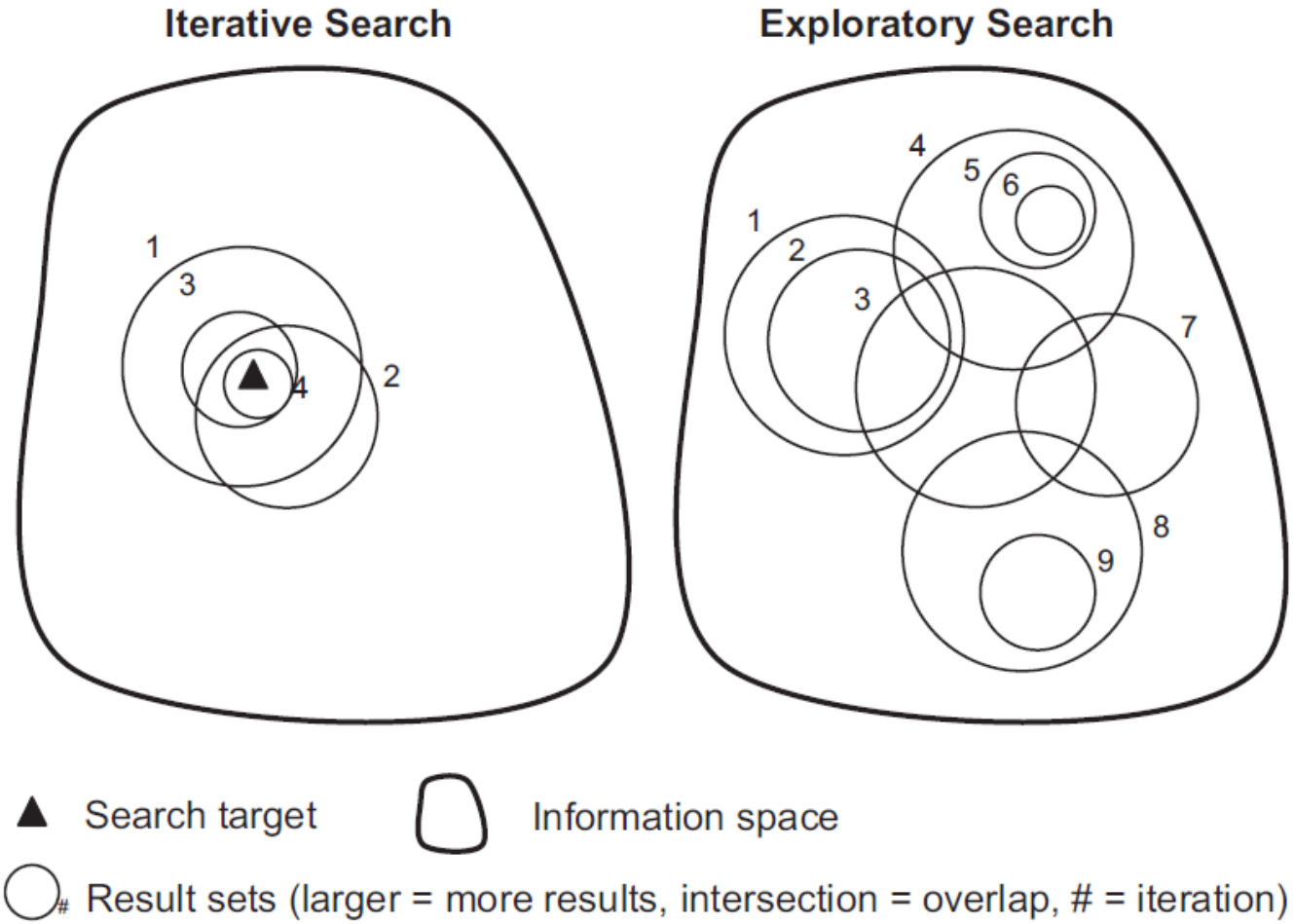
Концепция разведочного поиска (exploratory search)

- 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.

От ближнего чтения (close reading) к дальнему (distant reading)

Information Seeking Mantra [B.Shneiderman, 1996]

«Overview first, zoom and filter, details on demand»

Понятие *дальнего чтения* [Franco Moretti, 2005]

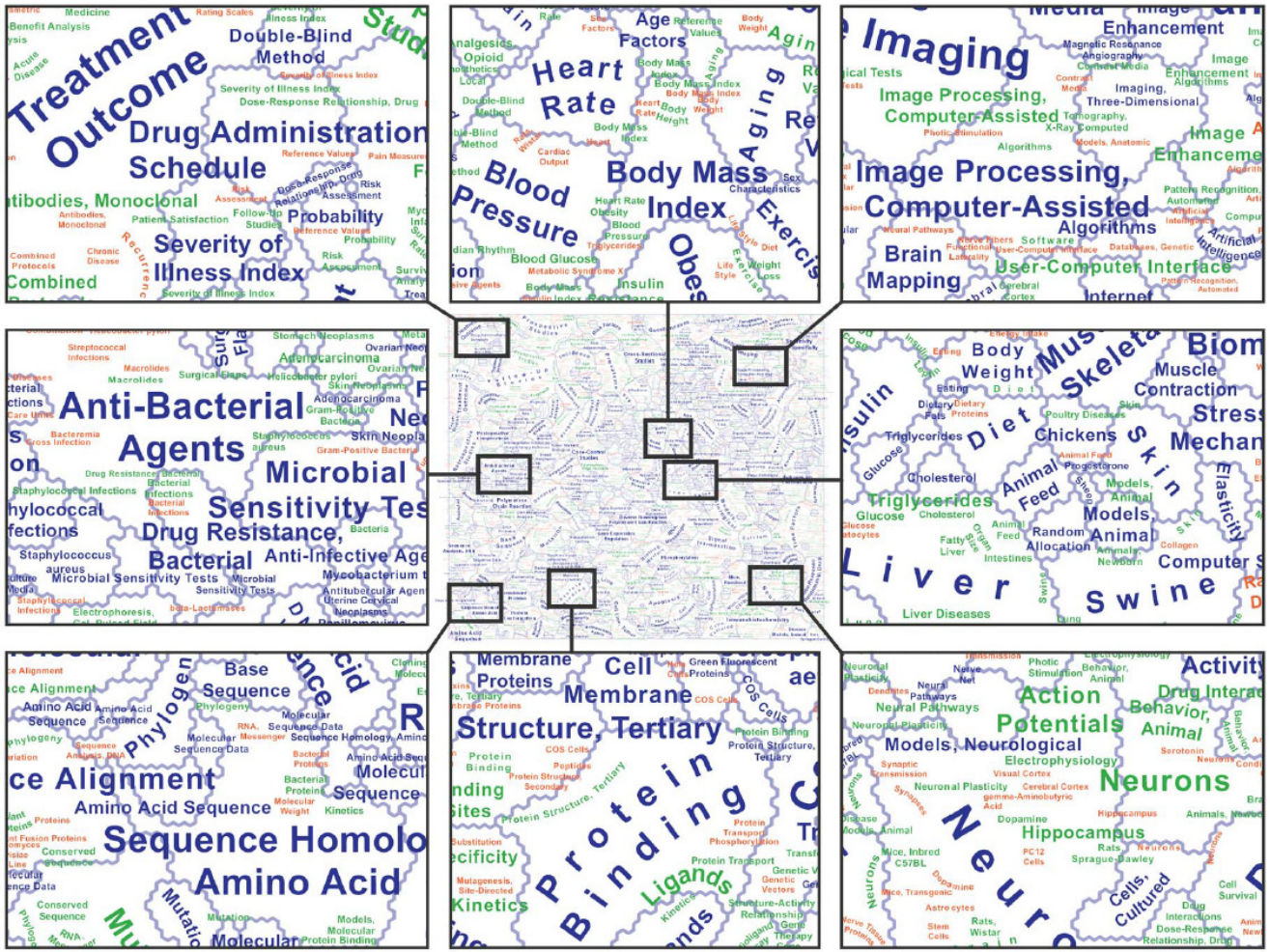
«*Distant reading* is not an obstacle but a specific form of knowledge: fewer elements, hence a sharper sense of their overall interconnection. Shapes, relations, structures. Forms. Models.»

B.Shneiderman. The Eyes Have It: A Task by Data Type Taxonomy for Information Visualizations. Visual Languages, 1996.

F.Moretti. Graphs, Maps, Trees: Abstract Models for a Literary History. 2005.

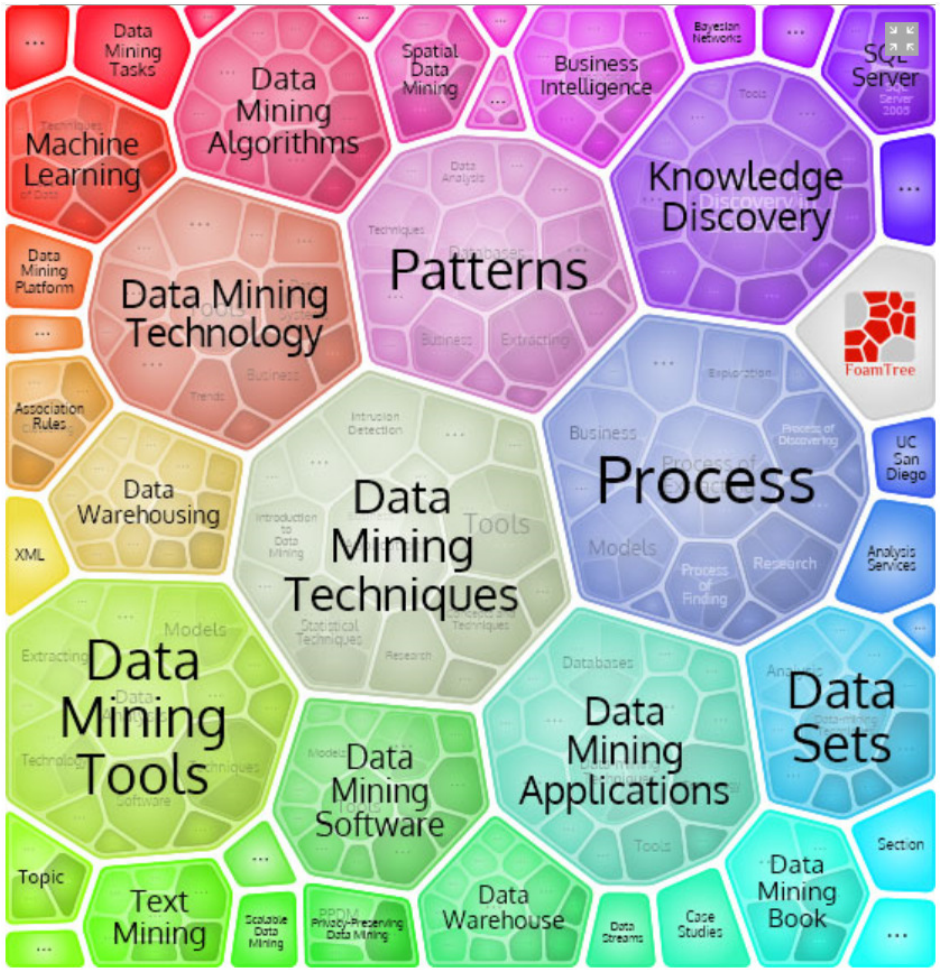
S.Janicke, G.Franzini, M.F.Cheema, G.Scheuermann. On Close and Distant Reading in Digital Humanities: A Survey and Future Challenges. EuroVis, 2015.

Пример карты медицинских знаний



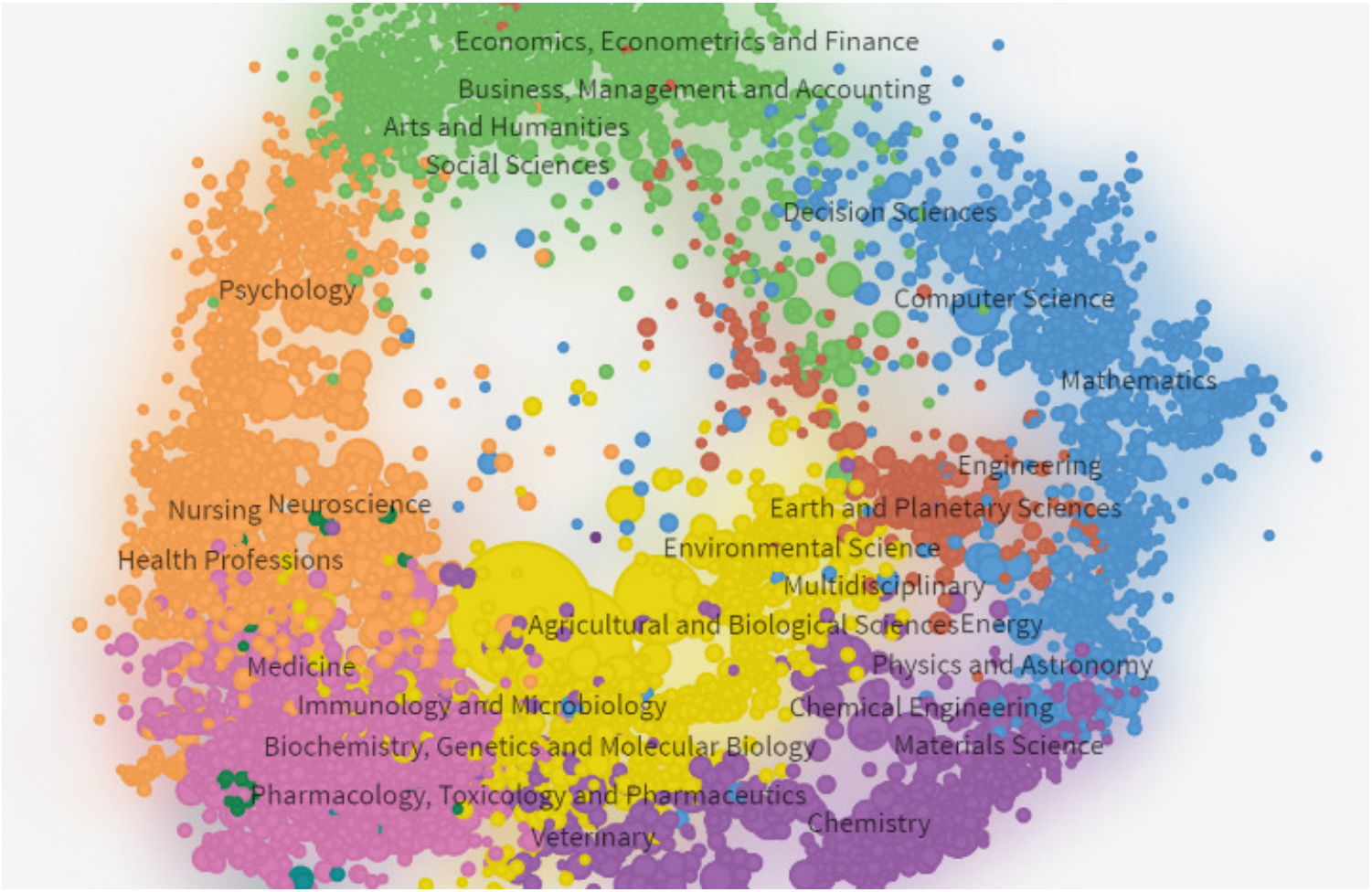
Skupin, Biberstine, Borner. Visualizing the Topical Structure of the Medical Sciences: A Self-Organizing Map Approach. PLoS ONE, 2013.

Пример иерархической карты области *Data Mining*



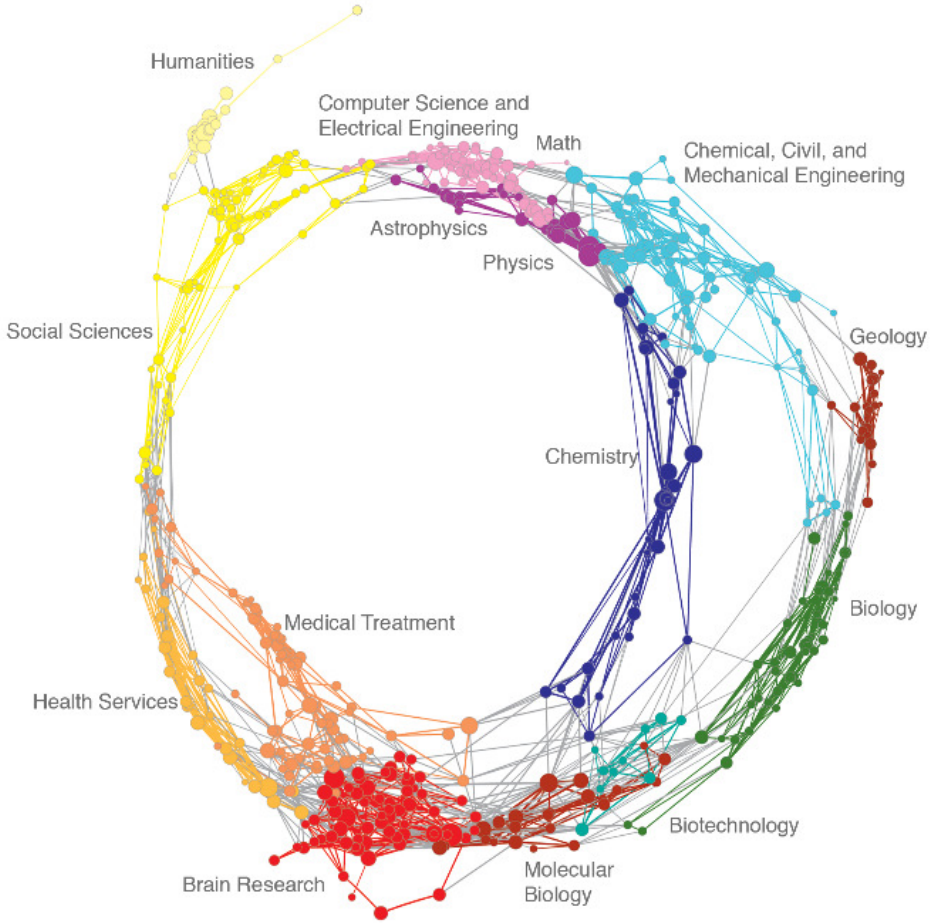
FoamTree: https://carrotsearch.com/foamtree

Пример карты науки



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

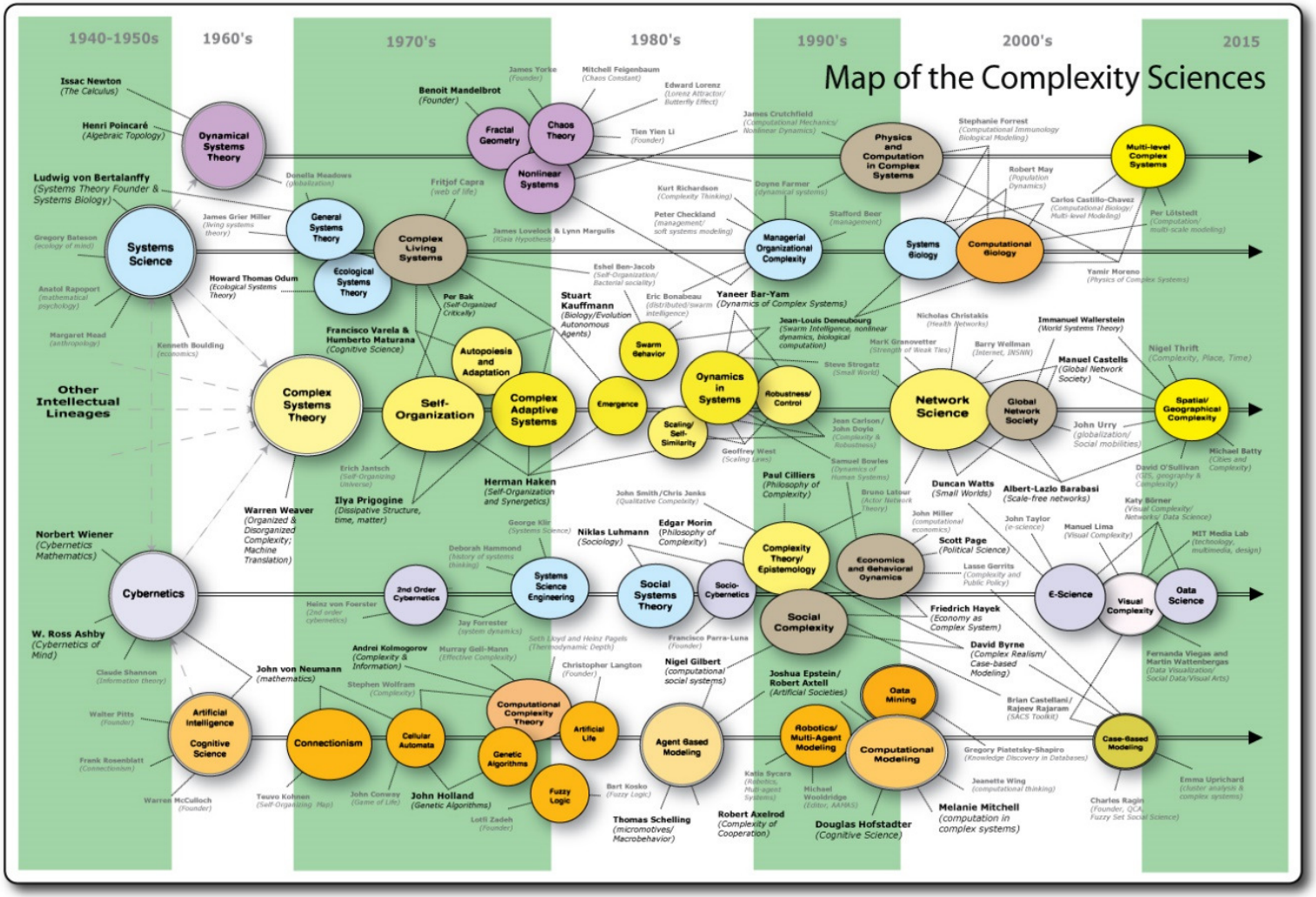
Ещё один пример карты науки



Недостатки: неинтерпретируемость осей, искажение сходства

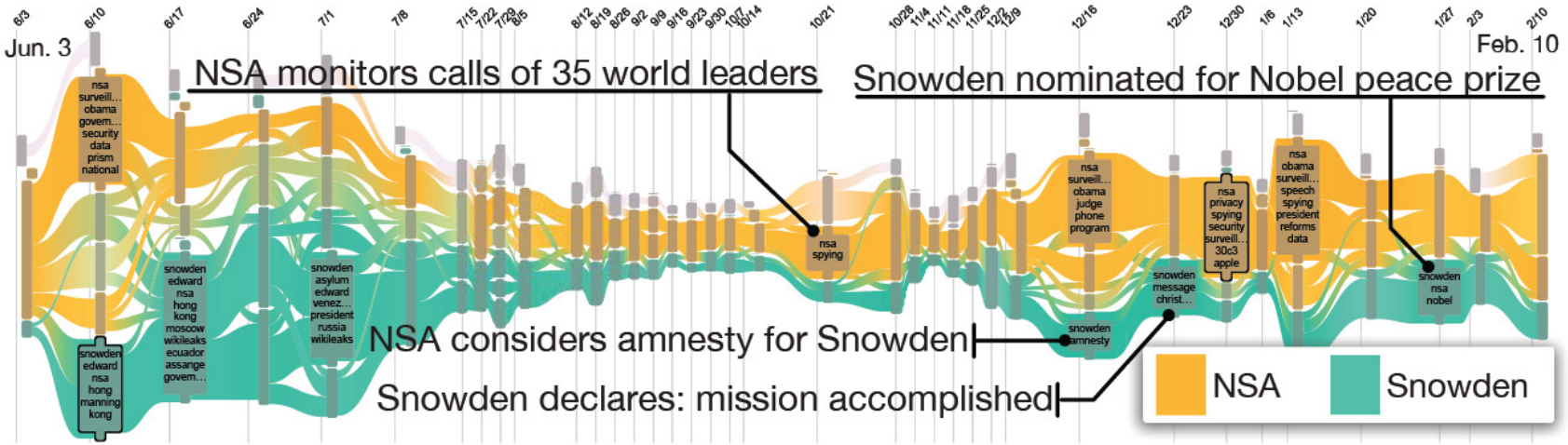
http://scimaps.org

Пример карты предметной области, построенной вручную



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

Динамика тем: эволюция предметной области



Эволюция выбранных тем иерархии. Данные Prism (2013/06/03–2014/02/09)

- эксперт задаёт сечение иерархии (дерева) тем,
- интерактивно выбирает подмножество тем и событий,
- генерирует отчёт.

Weiwei Cui, Shixia Liu, Zhuofeng Wu, Hao Wei. How hierarchical topics evolve in large text corpora. 2014.

Интерактивный обзор 330 средств визуализации текстов

http://textvis.lnu.se

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. получаем картину содержащихся в нём тем-подтем
3. и «дорожную карту» предметной области в целом

Представление результатов поиска (концепт)

- Двумерная карта в интерпретируемых осях тема–время
- Интерактивные возможности: zoom / filter / details
- Ось тем: гуманитарные $\rightarrow$ естественные $\rightarrow$ точные
- Темы делятся на подтемы иерархически

Технологические элементы разведочного поиска

По всем технологиям имеются готовые решения:

1. интернет-краулинг
2. фильтрация контента
3. тематическое моделирование — технология BigARTM
4. инвертированный индекс
5. ранжирование
6. визуализация
7. персонализация

Наша научная группа развивает теорию и технологии тематического моделирования как ключевой и наиболее наукоёмкий элемент разведочного поиска.

Что такое «тема»?

- *тема* — семантически однородный кластер текстов
- *тема* — специальная терминология предметной области
- *тема* — набор терминов (слов или словосочетаний), совместно часто встречающихся в документах
- тем много меньше, чем терминов и чем документов

Более формально,

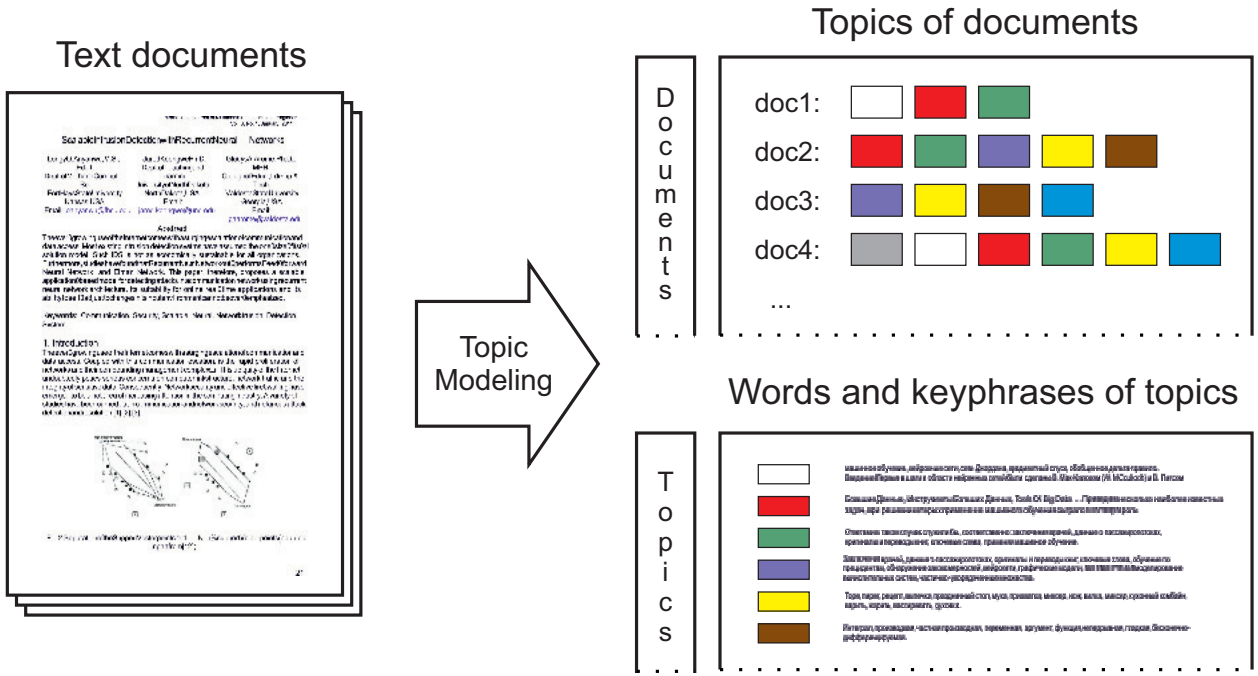
- *тема* — условное распределение на множестве терминов, $p(w|t)$ — вероятность (частота) термина w в теме t;
- *тематический профиль* документа — условное распределение $p(t|d)$ — вероятность (частота) темы t в документе d.

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

Что такое «тематическая модель»

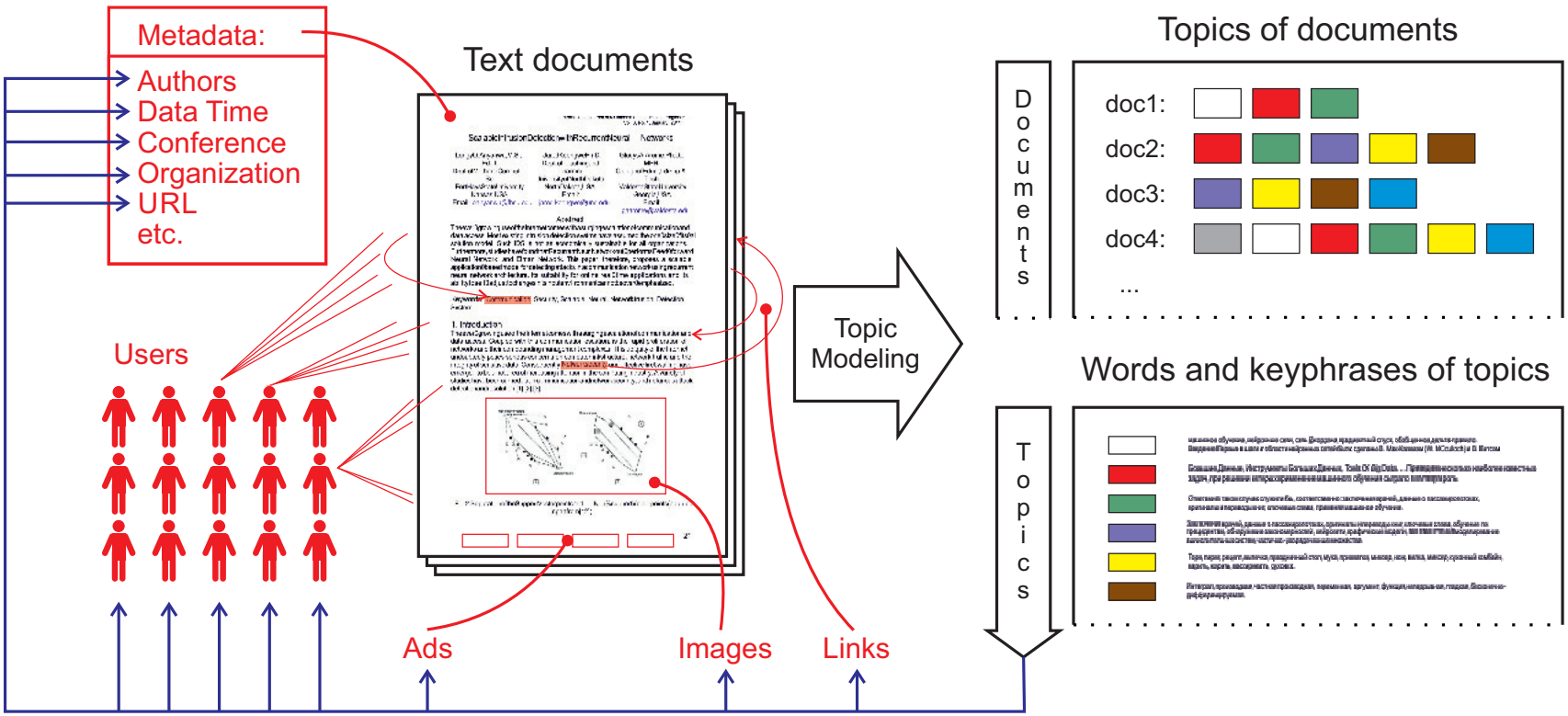
Вход: коллекция документов

Выход: тематика документов $p(t|d)$ и терминов $p(t|w)$:



Что такое «мультимодальная тематическая модель»

Выход: тематика документов $p(t|d)$ и терминов $p(t|w)$,
а также *модальностей*: $p(t|\text{автор})$, $p(t|\text{время})$, $p(t|\text{ссылка})$,
$p(t|\text{баннер})$, $p(t|\text{элемент изображения})$, $p(t|\text{пользователь})$,...



Пример тем. Мультиязычная модель Википедии

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

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

Пример тем. Мультиязычная модель Википедии

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

Асессор оценил 396 тем из 400 как хорошо интерпретируемые.

Биграммы радикально улучшают интерпретируемость тем

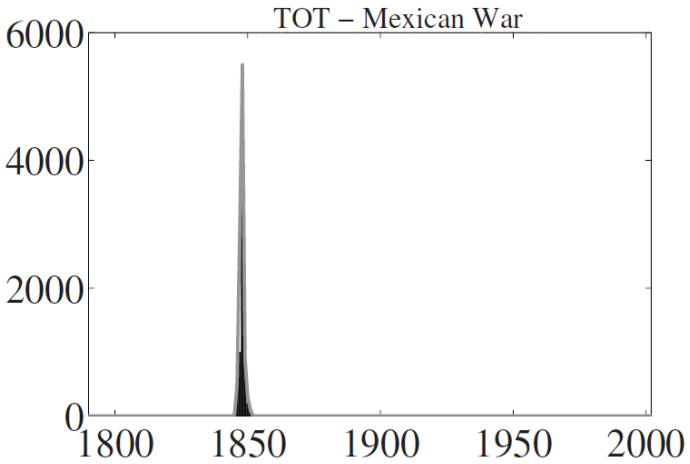
Коллекция 850 статей конференций ММРО, ИОИ на русском

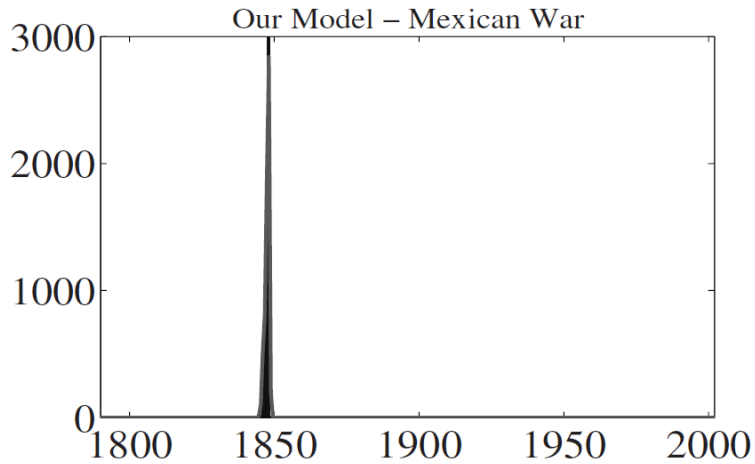
распознавание образов в биоинформатике		теория вычислительной сложности	
unigrams	bigrams	unigrams	bigrams
объект	задача распознавания	задача	разделять множества
задача	множество мотивов	множество	конечное множество
множество	система масок	подмножество	условие задачи
мотив	вторичная структура	условие	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

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

Совмещение динамической и n-граммной модели

По коллекции выступлений президентов США





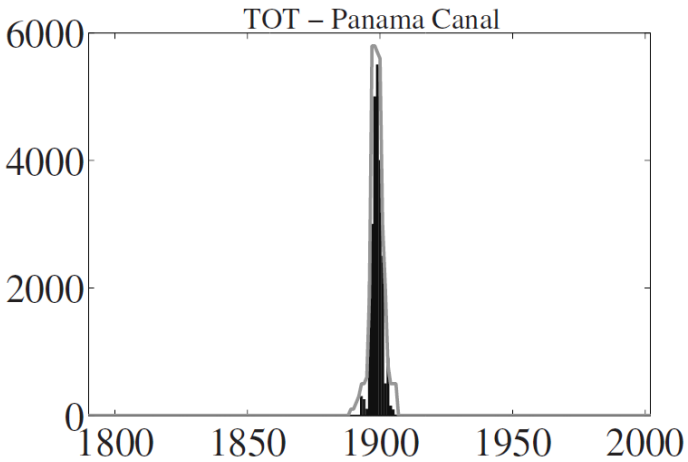
1. mexico	8. territory
2. texas	9. army
3. war	10. peace
4. mexican	11. act
5. united	12. policy
6. country	13. foreign
7. government	14. citizens

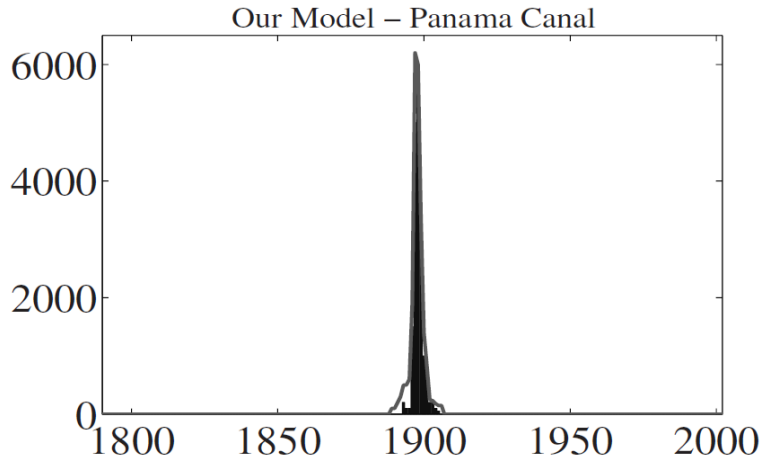
1. east bank	8. military
2. american coins	9. general herrera
3. mexican flag	10. foreign coin
4. separate independent	11. military usurper
5. american commonwealth	12. mexican treasury
6. mexican population	13. invaded texas
7. texan troops	14. veteran troops

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

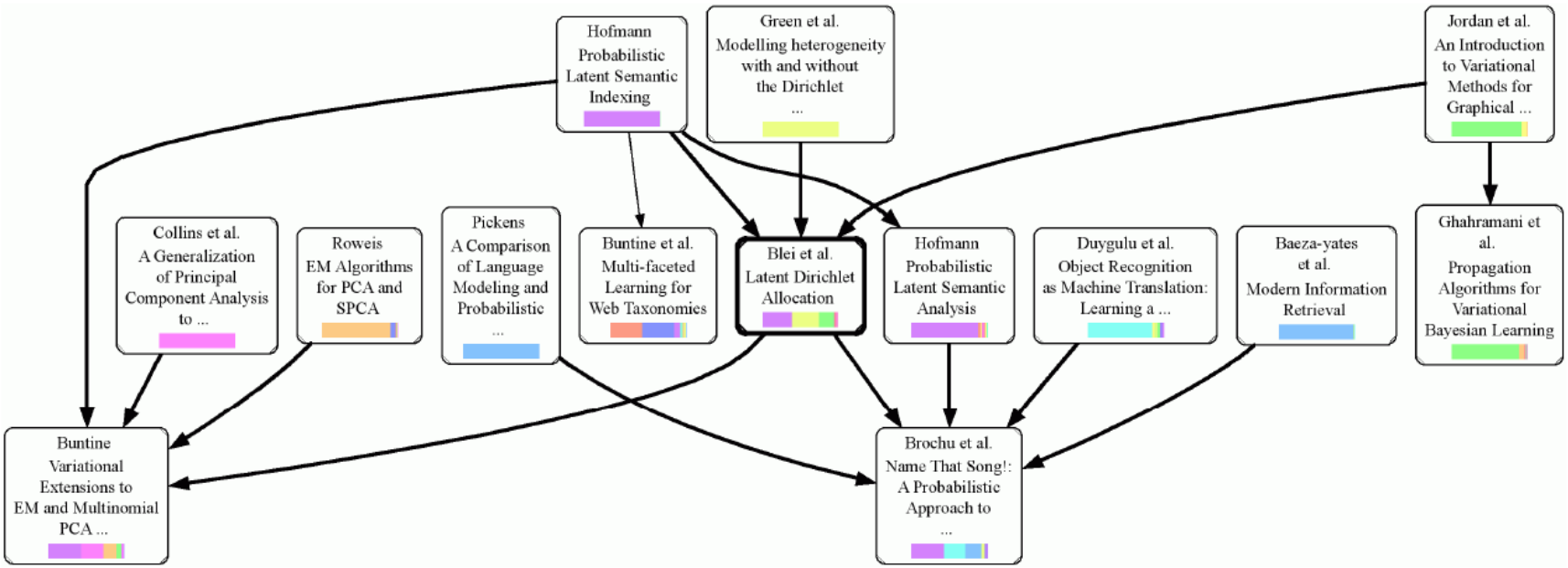
1. panama canal	8. united states senate
2. isthmian canal	9. french canal company
3. isthmus panama	10. caribbean sea
4. republic panama	11. panama canal bonds
5. united states government	12. panama
6. united states	13. american control
7. state panama	14. canal

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

Модели, учитывающие цитирования или гиперссылки

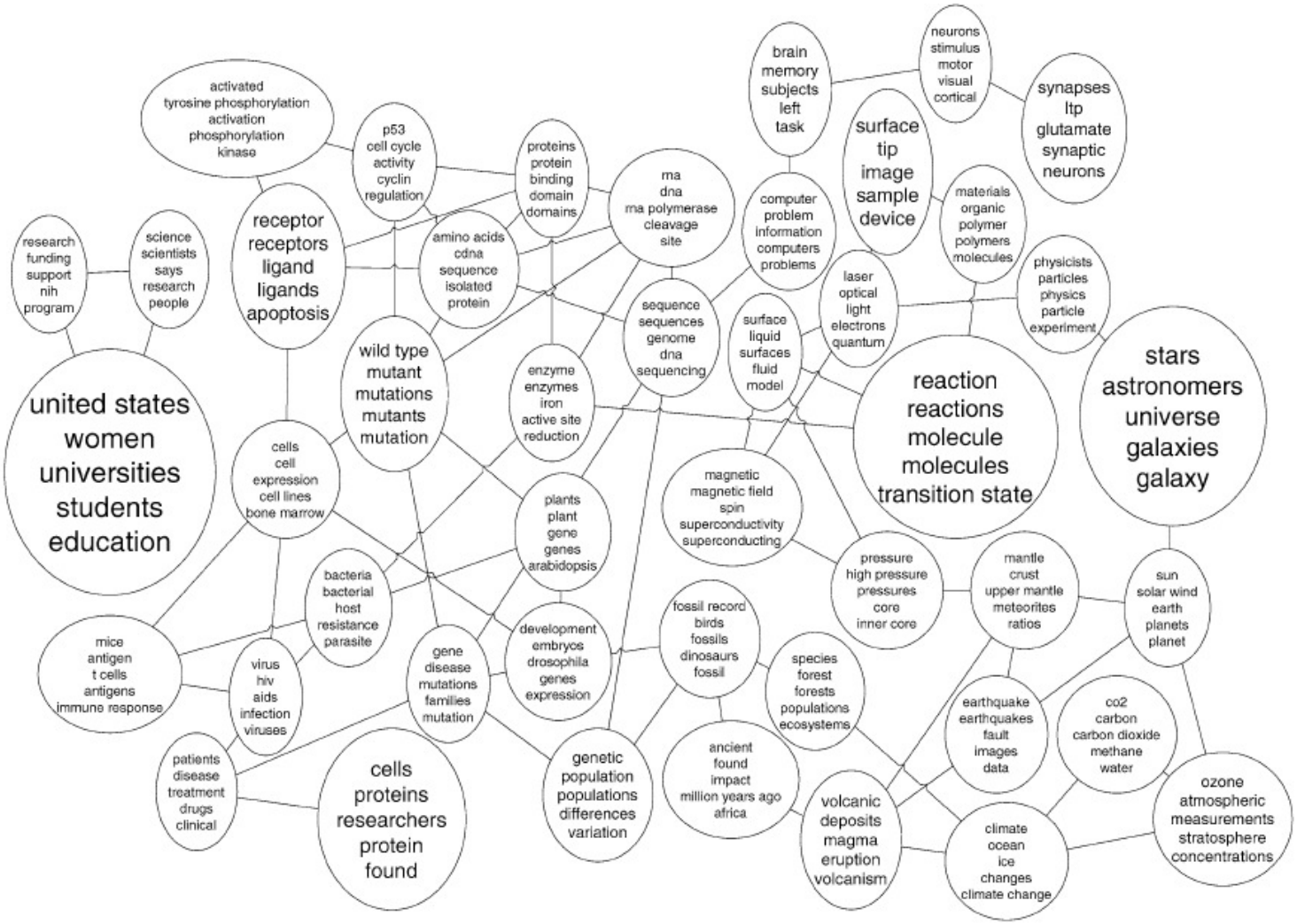
Учёт ссылок уточняет тематическую модель

Тематическая модель выявляет самые влиятельные ссылки



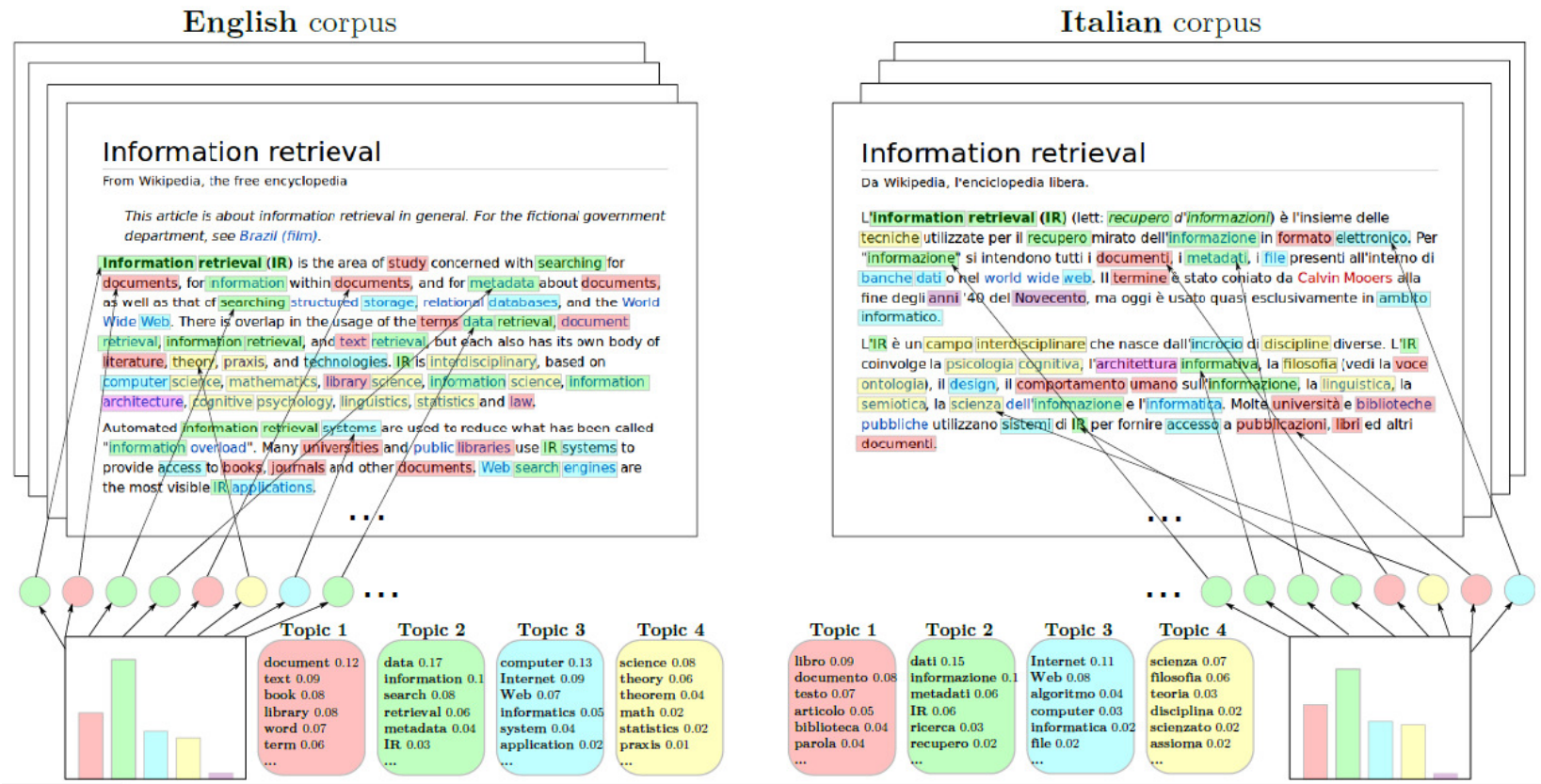
Laura Dietz, Steffen Bickel, Tobias Scheffer. Unsupervised prediction of citation influences. ICML-2007.

Выявление взаимосвязей между темами



D. Blei, J. Lafferty. A correlated topic model of Science. 2007.

Многоязычные модели параллельных коллекций



Неожиданное открытие: двуязычные словари не нужны!

I. Vulić, W. De Smet, J. Tang, M.-F. Moens. Probabilistic topic modeling in multilingual settings: a short overview of its methodology with applications. 2012.

Тематическая модель для разведочного поиска должна быть...

1. **Темпоральная**: отображение динамики развития тем
2. **Иерархическая**: систематизация областей знания
3. **Интерпретируемая**: каждая тема понятна для людей
4. **Мультиграммная**: выделение тематичных словосочетаний
5. **Мультимодальная**: авторы, связи, тэги, пользователи,...
6. **Мультиязычная**: кросс- и много-языковой поиск
7. **Разреженная**: для эффективности 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

Некоторые тематические модели

- 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) классификация
- BitermTM (2013) битермы в коротких документах
- mLDA (2013) метаданные с тремя и более модальностями
- WNTM (2014) локальные контексты слов

Библиотека тематического моделирования BigARTM

Ключевые возможности:

- Комбинирование требований, моделей, модальностей
- (благодаря теории аддитивной регуляризации, ARTM)
- Большие данные: коллекция не хранится в памяти
- Встроенная библиотека регуляризаторов и мер качества

Сообщество:

- Открытый код https://github.com/bigartm
- 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, Python, C++, C#

BigARTM: унификация разработки тематических моделей

На практике чаще всего используют устаревшую модель LDA.
Причина — байесовские модели приходится строить «с нуля».

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*

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* вычислена на тестовой выборке документов

Данные коллективного блога Хабрахабр.ру

Данные

- 132 157 статей
- Модальности:
  - 52 354 терминов (слов)
  - 524 авторов статей
  - 10 000 комментаторов (авторов комментариев к статьям)
  - 2546 тегов
  - 123 хаба (категории)

Предобработка текстов

- отброшены 5% наиболее частотных слов (общая лексика)
- удаление пунктуации
- нижний регистр, ё→е
- лемматизация pymorphy2

Разведочный поиск

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

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

Ранжируем документы коллекции $d \in D$ по убыванию косинусной меры близости документа d и запроса q.

Выдача тематического поиска — k первых документов.

Техническая реализация: *инвертированный индекс* для быстрого поиска документов d по каждой из тем t запроса

Методика оценивания качества разведочного поиска

Поисковый запрос

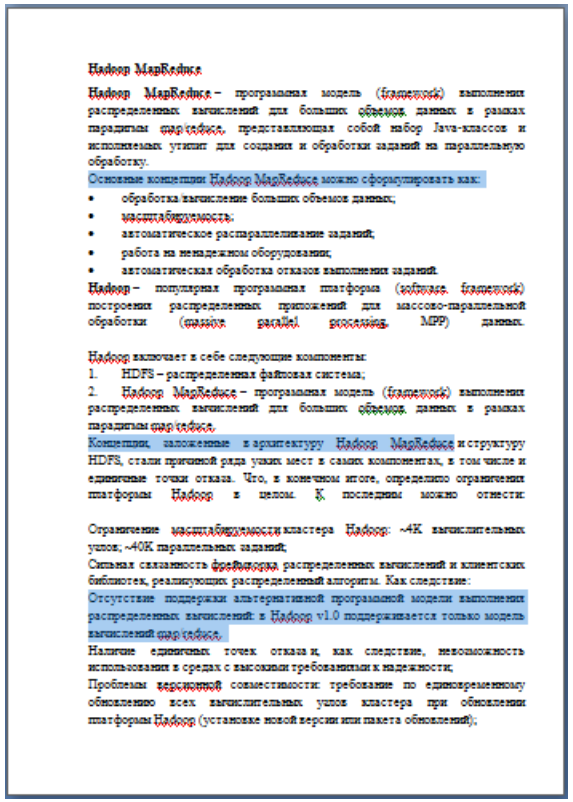
набор ключевых слов или фрагментов текста, около одной страницы A4

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



Hadoop MapReduce

Hadoop MapReduce – программная модель (framework) выполнения распределенных вычислений для больших объемов данных в рамках парадигмы map/reduce, представляющая собой набор Java-классов и исполняемых утилит для создания и обработки заданий на параллельную обработку.

Основные концепции Hadoop MapReduce можно сформулировать как:

- обработка/вычисление больших объемов данных;
- масштабируемость;
- автоматическое распараллеливание заданий;
- работа на ненадежном оборудовании;
- автоматическая обработка отказов выполнения заданий.

Hadoop – популярная программная платформа (software framework) построения распределенных приложений для массово-параллельной обработки (massive parallel processing, MPP) данных.

Hadoop включает в себе следующие компоненты:

1. HDFS – распределенная файловая система;
2. Hadoop MapReduce – программная модель (framework) выполнения распределенных вычислений для больших объемов данных в рамках парадигмы map/reduce.

Концепции, заложенные в архитектуру Hadoop MapReduce и структуру HDFS, стали причиной ряда узких мест в самих компонентах, в том числе и единичные точки отказа. Что, в конечном итоге, определило ограничения платформы Hadoop в целом. К последним можно отнести:

Ограничение масштабируемости кластера Hadoop: ~4K вычислительных узлов; ~40K параллельных заданий;

Сильная связанность фреймворка распределенных вычислений и клиентских библиотек, реализующих распределенный алгоритм. Как следствие:

Отсутствие поддержки альтернативной программной модели выполнения распределенных вычислений: в Hadoop v1.0 поддерживается только модель вычислений map/reduce.

Наличие единичных точек отказа и, как следствие, невозможность использования в средах с высокими требованиями к надежности;

Проблемы версионной совместимости: требование по единовременному обновлению всех вычислительных узлов кластера при обновлении платформы Hadoop (установке новой версии или пакета обновлений);

Пример запроса для разведочного поиска

Пример: фрагмент запроса «Система IBM Watson»

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

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

... ...

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

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

Тематика запросов разведочного поиска

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

Оценки качества поиска

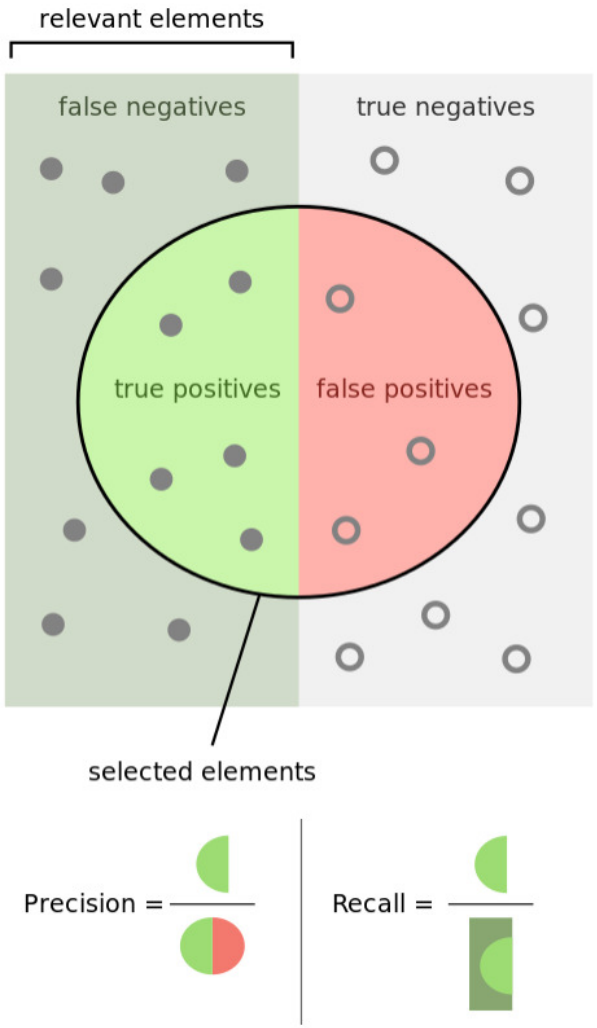
Precision — доля релевантных среди найденных
Recall — доля найденных среди релевантных

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

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

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

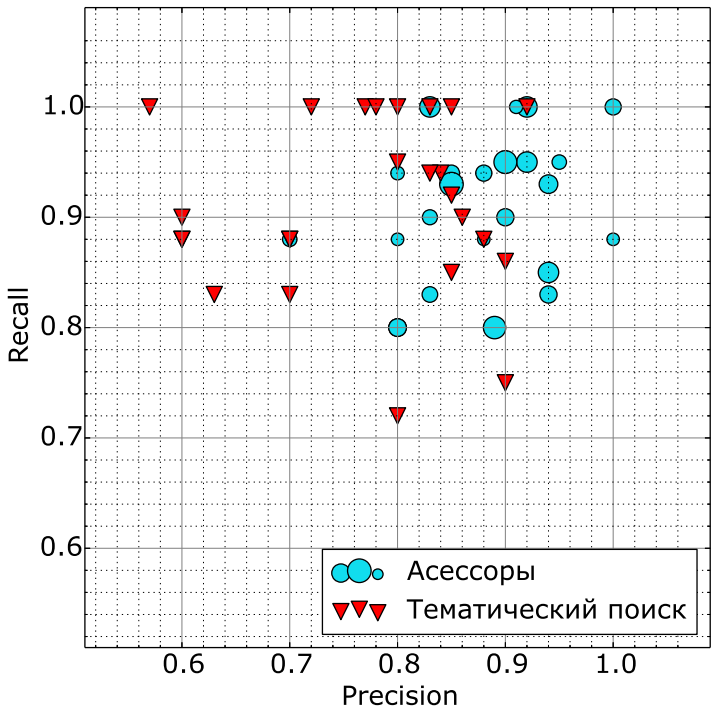
TP (true positive) — найденные релевантные
FP (false positive) — найденные нерелевантные
FN (false negative) — ненайденные релевантные



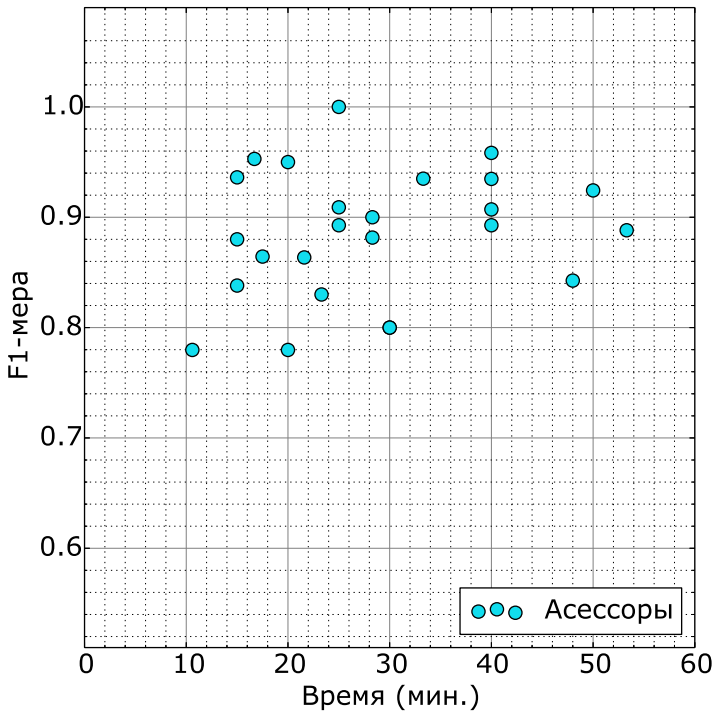
Результаты измерения точности и полноты по запросам

25 запросов, 3 асессора на запрос

точность и полнота поиска



время и F_1-мера (асессоры)



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

Выбор модальностей по критериям точности и полноты

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

	acессоры	C	K	TX	CT	CX	CTX	все
Precision@5	0.82	0.63	0.54	0.59	0.74	0.73	0.73	**0.74**
Precision@10	0.87	0.67	0.56	0.58	0.77	0.74	0.75	**0.77**
Precision@15	0.86	0.65	0.53	0.55	0.67	0.67	0.68	**0.68**
Precision@20	0.85	0.64	0.53	0.54	0.66	0.67	0.68	**0.68**
Recall@5	0.78	0.77	0.63	0.69	0.82	0.81	0.82	**0.82**
Recall@10	0.84	0.79	0.64	0.71	0.88	0.82	0.87	**0.88**
Recall@15	0.88	0.82	0.67	0.73	0.90	0.84	0.89	**0.90**
Recall@20	0.88	0.85	0.68	0.74	0.91	0.85	0.89	**0.91**

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

Выбор числа тем по критериям точности и полноты

Теперь используем все 5 модальностей, меняем число тем $|T|$

	асессоры	100	200	300	400	500
Precision@5	0.82	0.61	**0.74**	0.71	0.69	0.59
Precision@10	0.87	0.65	**0.77**	0.72	0.67	0.61
Precision@15	0.86	0.67	**0.68**	0.67	0.65	0.62
Precision@20	0.85	0.64	**0.68**	0.67	0.64	0.60
Recall@5	0.78	0.62	**0.82**	0.80	0.72	0.63
Recall@10	0.84	0.63	**0.88**	0.81	0.75	0.64
Recall@15	0.88	0.67	**0.90**	0.82	0.77	0.67
Recall@20	0.88	0.69	**0.91**	0.85	0.77	0.68

- Наилучшее качество поиска — при 200 темах
- Тематический поиск превосходит асессоров по полноте

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

Резюме: возможности тематического моделирования

- Разведочный поиск по длинным запросам
- Автоматическая рубрикация больших коллекций
- Тематический мониторинг входного потока документов
- Тематические рекомендации пользователям
- Формирование данных визуальных карт
- Обработка больших текстовых коллекций
- Учёт метаинформации и гетерогенных данных

http://bigartm.org

Литература

- *K.Vorontsov.* Additive regularization for topic models of text collections. 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. (на рецензии)