Вероятностные тематические модели коллекций текстовых документов

К. В. Воронцов
vokov@forecsys.ru

(Вычислительный Центр им. А. А. Дородницына РАН)

Научный семинар ВИНИТИ РАН • 23 апреля 2013

Содержание

Тема как статистическое явление в коллекциях документов

Тема — это набор терминов, неслучайно часто совместно встречающихся в относительно узком подмножестве документов.

Дано:

W — словарь, множество слов (терминов)

D — множество (коллекция, корпус) текстовых документов

n_{dw} — сколько раз термин $w \in W$ встретился в документе $d \in D$

Постановки задач:

— найти, какими терминами определяется каждая тема

— найти, к каким темам относится каждый документ

— определить число статистически различимых тем

— восстановить иерархию тем

— построить динамику развития тем во времени

— найти тематику связанных с документами объектов

Цели тематического моделирования (topic modeling)

- Тематический поиск документов и объектов по тексту любой длины или по любому объекту
- Категоризация, классификация, аннотирование, суммаризация текстовых документов

Типичные приложения:

- Поиск научной информации
- Поиск экспертов (expert search), рецензентов, проектов
- Выявление трендов и фронта исследований
- Анализ и агрегирование новостных потоков
- Рубрикация документов, изображений, видео, музыки
- Рекомендательные сервисы (коллаборативная фильтрация)
- Аннотация генома и другие задачи биоинформатики

Этапы предварительной обработки текстов

- Удаление переносов, чисел, колонтитулов, таблиц, остатков формул, оглавлений и т. д.
- Удаление опечаток и ошибок сканирования
- Приведение всех слов к нормальной форме (лемматизация или стемминг)
- Удаление общеупотребительных слов (стоп-слов)
- Удаление слишком редких специфических слов
- Выделение терминов (term extraction) и/или выделение словосочетаний (key phrase extraction) (сводятся к задачам классификации или ранжирования)
- Распознавание класса документа (научный? реферат? художественный? публицистика?)
- Выделение метаописания: название, авторы, год и т. д.
- Выделение библиографических ссылок

Формализация постановки задачи

Гипотезы о статистической природе текстов:

1. коллекция — это цепочка троек $\langle \text{документ}_i, \text{термин}_i, \text{тема}_i \rangle$
2. гипотезы «мешка слов» и «мешка документов»
3. тема t — это латентная переменная, и мы хотим найти
   $p(w|t) = n_{wt}/n_t$ — распределение терминов w в темах t,
   $p(t|d) = n_{td}/n_d$ — распределение тем t в документах d
4. гипотеза разреженности: среди $p(w|t)$, $p(t|d)$ много нулей
5. гипотеза условной независимости: $p(w|d,t) = p(w|t)$;

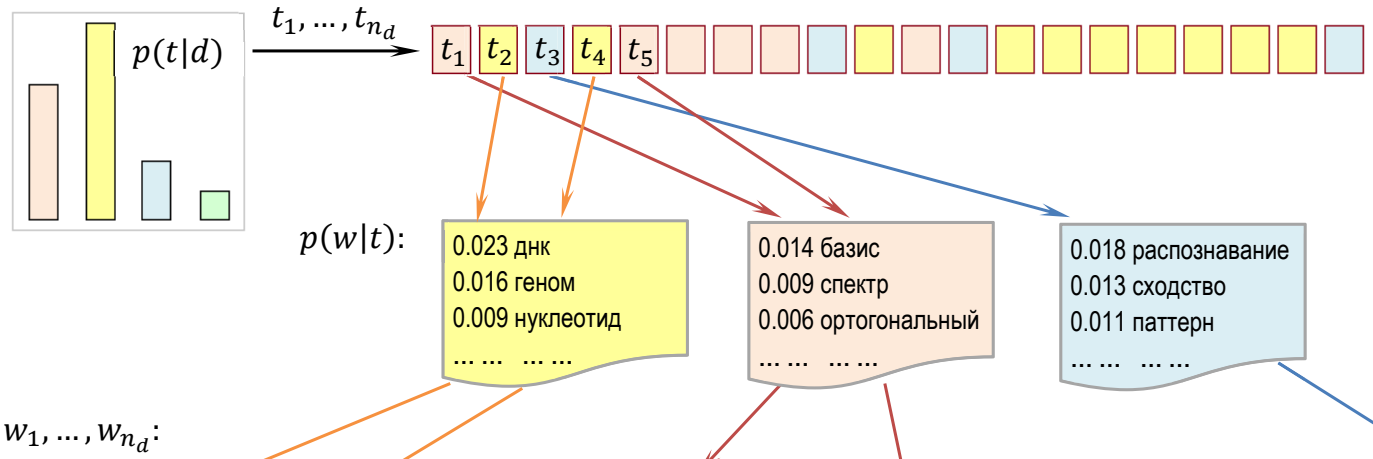
Вероятностная модель порождения документа d:

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

Задача: зная $p(w|d) = n_{dw}/n_d$, найти $p(w|t)$, $p(t|d)$.

Вероятностная модель порождения документа d

Вероятностная тематическая модель: $p(w|d) = \sum_{t \in T} p(w|t)p(t|d)$



Разработан спектрально-аналитический подход к выявлению размытых протяженных повторов в геномных последовательностях. Метод основан на разномасштабном оценивании сходства нуклеотидных последовательностей в пространстве коэффициентов разложения фрагментов кривых GC- и GA-содержания по классическим ортогональным базисам. Найдены условия оптимальной аппроксимации, обеспечивающие автоматическое распознавание повторов различных видов (прямых и инвертированных, а также тандемных) на спектральной матрице сходства. Метод одинаково хорошо работает на разных масштабах данных. Он позволяет выявлять следы сегментных дупликаций и мегасателлитные участки в геноме, районы синтении при сравнении пары геномов. Его можно использовать для детального изучения фрагментов хромосом (поиска размытых участков с умеренной длиной повторяющегося паттерна).

Модель PLSA — Probabilistic Latent Semantic Analysis (1999)

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

E-шаг: условные вероятности тем $p(t|d,w)$ по формуле Байеса:

$$p(t|d,w) = \frac{p(w|t)p(t|d)}{\sum_s p(w|s)p(s|d)}.$$

M-шаг: оценить $n_{dwt} = n_{dw}p(t|d,w)$ — сколько раз слово w в документе d относится к теме t;
затем вычислить частотные оценки условных вероятностей:

$$p(w|t) = \frac{n_{wt}}{n_t}, \qquad n_{wt} = \sum_{d\in D} n_{dwt}, \qquad n_t = \sum_{w\in W} n_{wt};$$

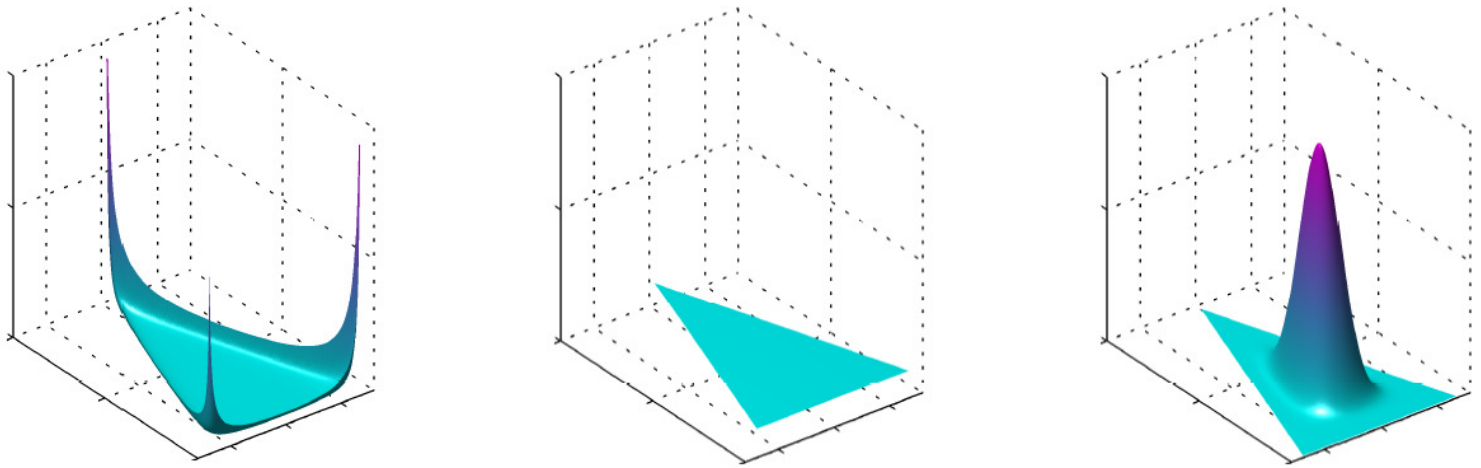
$$p(t|d) = \frac{n_{dt}}{n_d}, \qquad n_{dt} = \sum_{w\in d} n_{dwt}, \qquad n_d = \sum_{t\in T} n_{dt}.$$

Thomas Hofmann. Probabilistic latent semantic indexing // ACM SIGIR, Berkeley, California, USA, 1999. — Pp. 50–57.

Модель LDA — Latent Dirichlet Allocation (D.Blei, 2003)

Модель латентного размещения Дирихле предполагает, что распределения $p(w|t)$, $p(t|d)$ — это случайные векторы, порождаемые двумя априорными распределениями Дирихле.

Пример. $p(t|d) \sim \text{Dir}(\alpha)$ в пространстве трёх тем, $|T| = 3$:



$\alpha_1 = \alpha_2 = \alpha_3 = 0.1$ $\alpha_1 = \alpha_2 = \alpha_3 = 1$ $\alpha_1 = \alpha_2 = \alpha_3 = 10$

David Blei, Andrew Ng, Michael Jordan. Latent Dirichlet allocation. Journal of Machine Learning Research, 2003. — No. 3. — Pp. 993–1022.

Обобщения и модификации тематических моделей

- Hierarchical models — строят иерархическую структуру тем
- Temporal models — учитывают динамику публикаций
- Author-topic models — оценивают распределение авторов $p(a|w,d)$ для каждого слова документа
- Entity-topic models — оценивают тематику объектов (веществ, животных, генов, стран, фирм, изделий,...)
- Multilingual topic models — многоязыковые модели
- Модели, учитывающие связь слов внутри документа
- Модели связей между документами (ссылки, цитирование)

Ali Daud, Juanzi Li, Lizhu Zhou, Faqir Muhammad.
Knowledge discovery through directed probabilistic topic models: a survey.
Frontiers of Computer Science in China, Vol. 4, No. 2., 2010, Pp. 280–301.
(русский перевод на www.MachineLearning.ru)

Topic Modeling Bibliography:
`http://www.cs.princeton.edu/~mimno/topics.html`

Иерархические тематические модели

Для выявления иерархии тем используется модель HDP — иерархический процесс Дирихле, обобщение модели LDA.

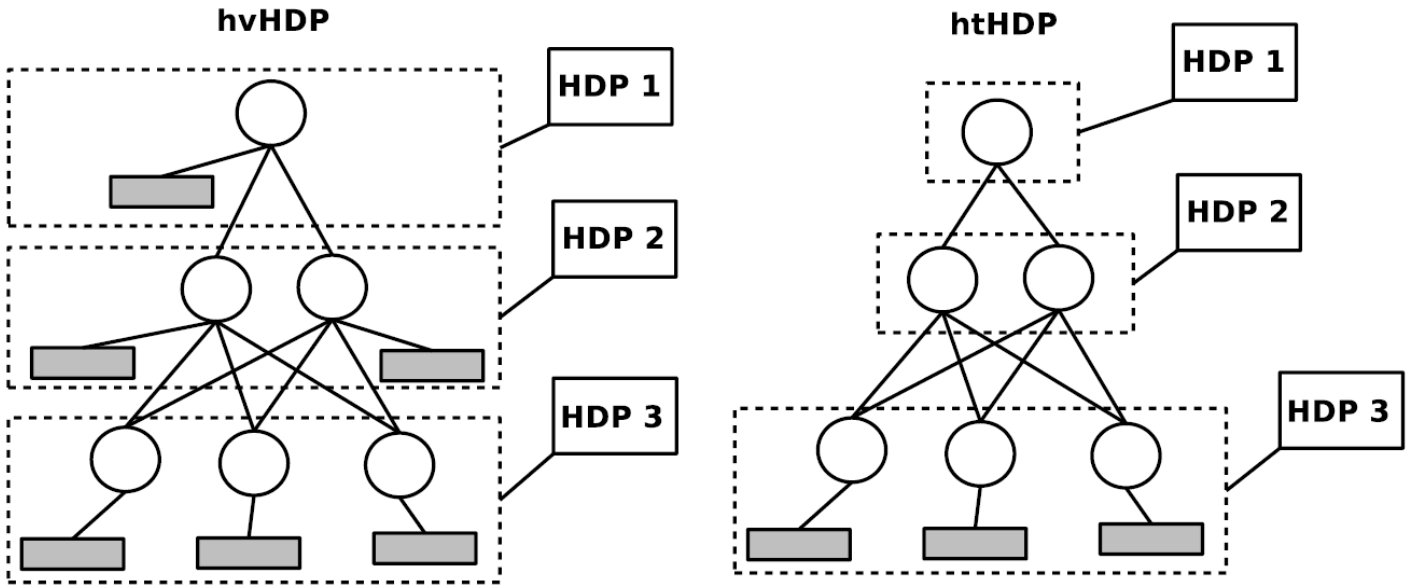
Задача построения иерархии и задача оценивания её качества признаются открытыми научными проблемами.

- "Despite recent activity in the field of HPTMs, determining the hierarchical model that best fits a given data set, in terms of the structure and size of the learned hierarchy, still remains a challenging task and an open issue."
- "The evaluation of topic models is also an open issue."

E. Zavitsanos, G. Paliouras, G. A. Vouros. Non-Parametric Estimation of Topic Hierarchies from Texts with Hierarchical Dirichlet Processes // Journal of Machine Learning Research 12 (2011) 2749-2775.

Две восходящие стратегии построения иерархии

- hvHDP: внутренние вершины — темы, имеющие $p(w|t)$
- htHDP: внутренние вершины — кластеры тем



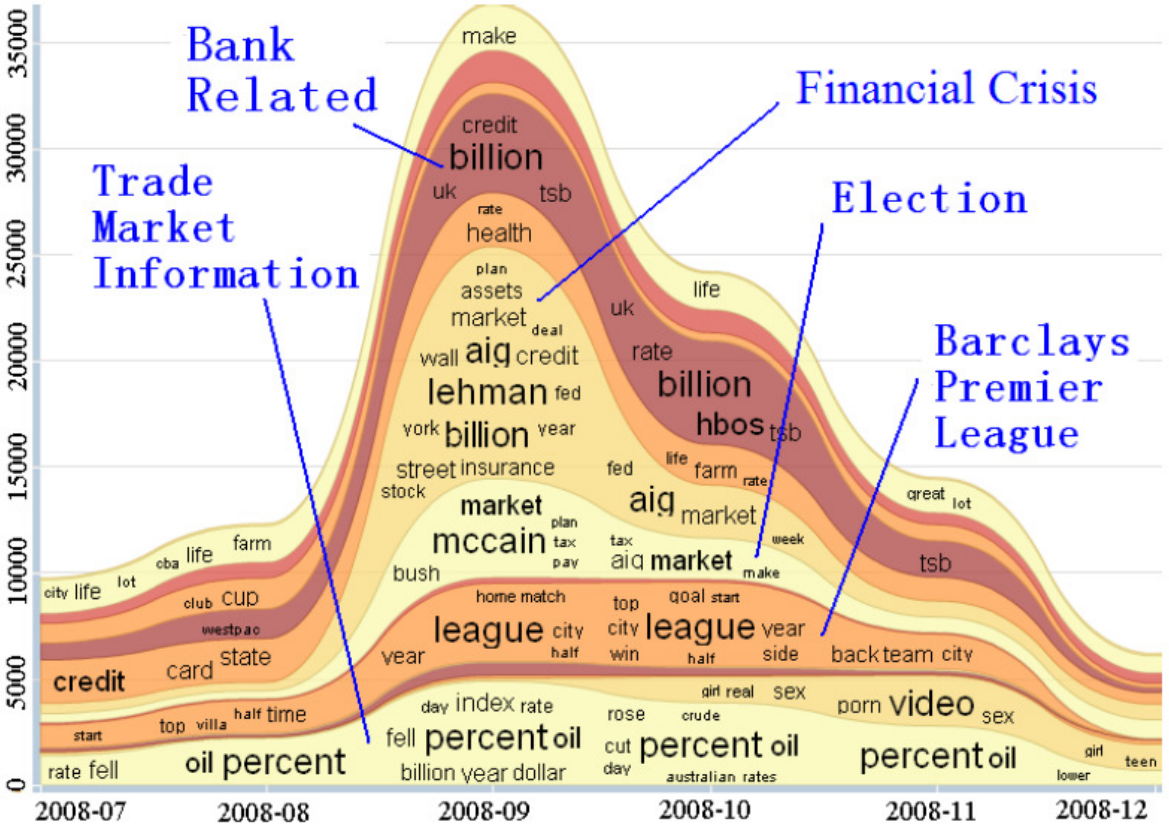
E. Zavitsanos, G. Paliouras, G. A. Vouros. Non-Parametric Estimation of Topic Hierarchies from Texts with Hierarchical Dirichlet Processes // Journal of Machine Learning Research 12 (2011) 2749-2775.

Динамические модели (temporal topic models)

Основные предположения:

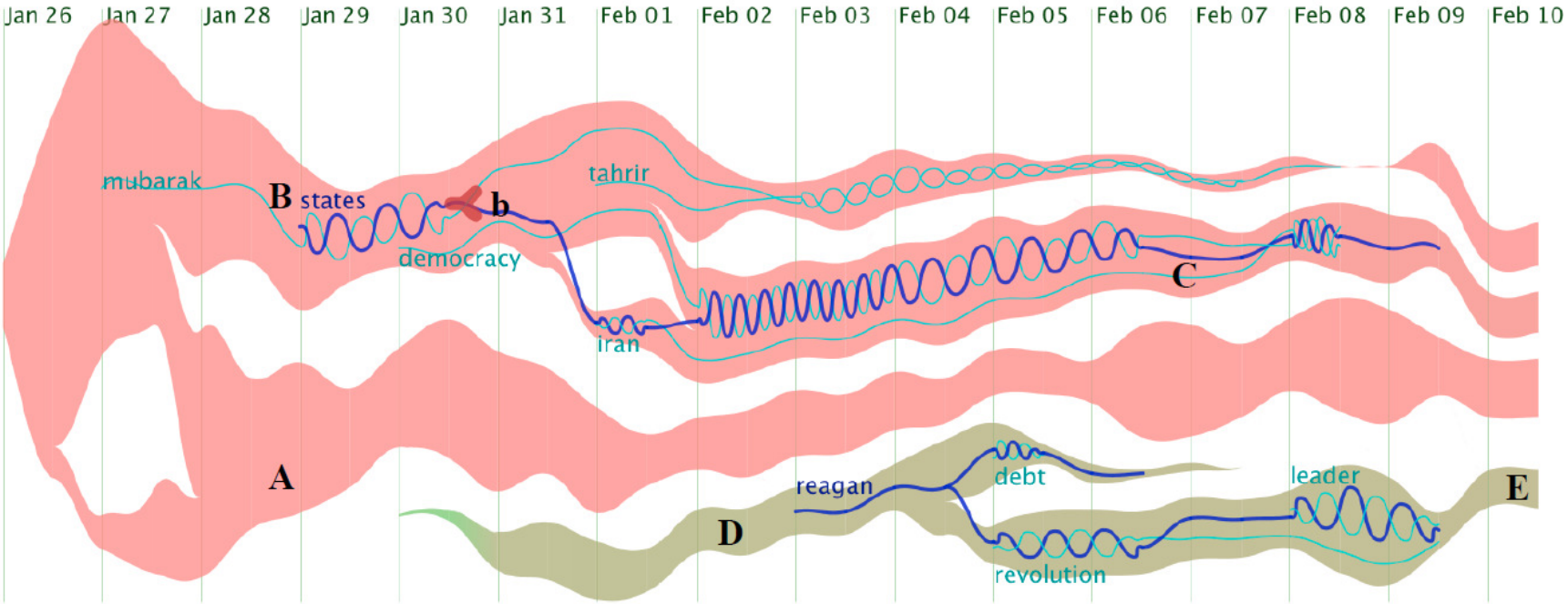
- Документы имеют привязку к моменту времени.
- Распределения $p(w|t,y)$ зависят от времени y.
- Резкие изменения происходят редко.
- Темы могут
  — появляться;
  — исчезать;
  — сливаться;
  — расщепляться.
- Для обнаружения этих событий используются статистические критерии.

Пример динамической модели



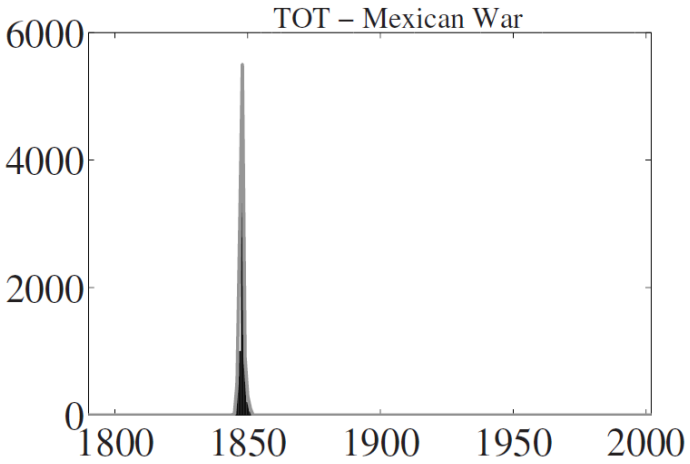
Jianwen Zhang, Yangqiu Song, Changshui Zhang, Shixia Liu Evolutionary Hierarchical Dirichlet Processes for Multiple Correlated Time-varying Corpora // KDD'10, July 25–28, 2010.

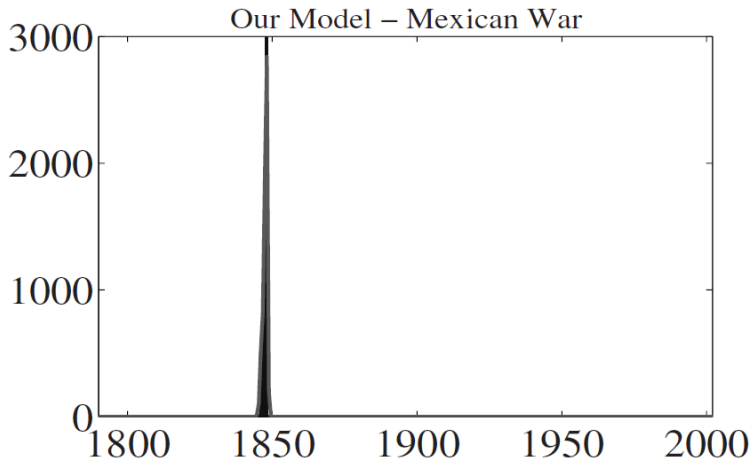
Ещё пример динамической модели



Weiwei Cui, Shixia Liu, Li Tan, Conglei Shi, Yangqiu Song, Zekai J. Gao, Xin Tong, Huamin Qu TextFlow: Towards Better Understanding of Evolving Topics in Text // IEEE Transactions On Visualization And Computer Graphics, Vol. 17, No. 12, December 2011.

Совмещение динамической и n-граммной модели



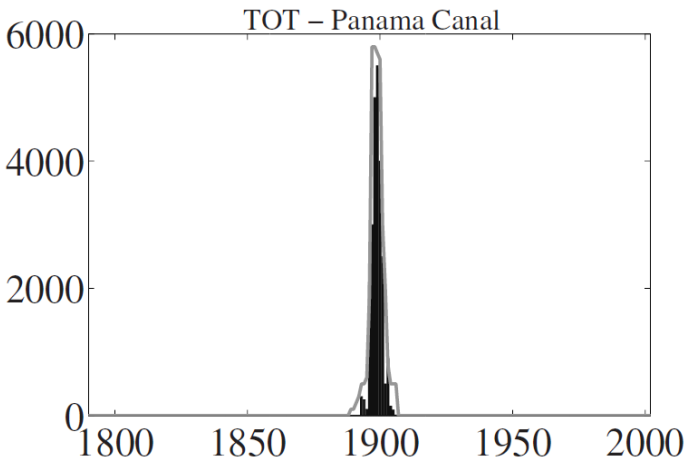


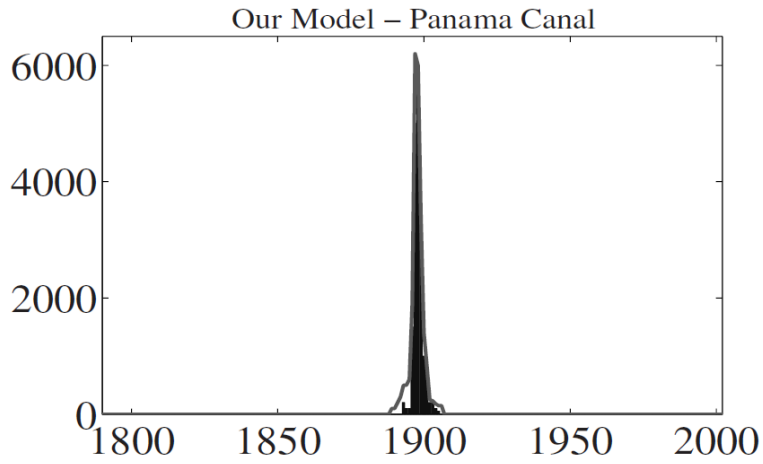
1. mexico	8. territory
2. texas	9. army
3. war	10. peace
4. mexican	11. act
5. united	12. policy
6. country	13. foreign
7. government	14. citizens

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 // 35'th ECIR 2013, Moscow, March 24–27. — pp. 292–304.

Совмещение динамической и n-граммной модели



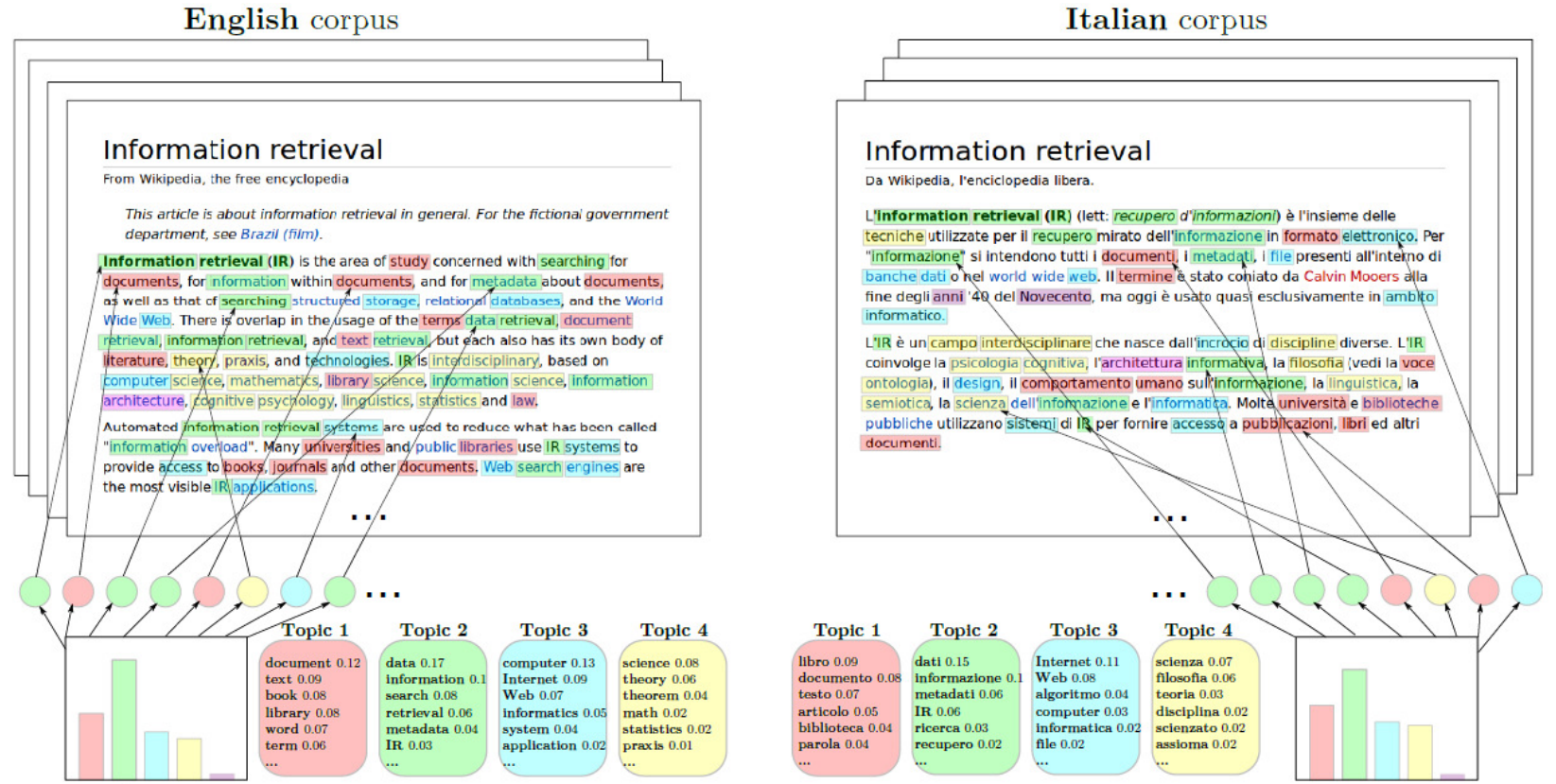


1. government	8. spanish
2. cuba	9. island
3. islands	10. act
4. international	11. commission
5. powers	12. officers
6. gold	13. spain
7. action	14. rico

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 // 35'th ECIR 2013, Moscow, March 24–27. — pp. 292–304.

Многоязычные модели



I. Vulić, W. De Smet, J. Tang, M.-F. Moens. Probabilistic topic modeling in multilingual settings: a short overview of its methodology with applications // NIPS, 7–8 December 2012. — Pp. 1–11.

Многоязычные модели

Основные выводы:

- достаточно приравнять $p(t|d)$ параллельных текстов
- достаточно иметь относительно небольшой параллельный корпус для каждой пары языков
- достаточно выравнивания на уровне документов
- выравнивание на уровне предложений или абзацев не обязательно, но улучшает качество многоязычного поиска
- наличие словаря «много-ко-многим» не обязательно, но улучшает качество многоязычного поиска
- нет необходимости применять «тяжёлые» методы машинного перевода

Оценивание качества тематических моделей

Внутренние критерии:

- правдоподобие/перплексия на обучении/контроле
- тест условной независимости (для отдельных тем)
- устойчивость

Внешние критерии:

- качество поиска документов по аннотации или фрагменту
- качество категоризации документов
- интерпретируемость тем

J. Chang, J. Boyd-Graber, S. Gerrish, C. Wang, D. Blei. Reading Tea Leaves: How Humans Interpret Topic Models // NIPS, 2009. — Pp. 288–296.

D. Mimno, D. Blei, Bayesian checking for topic models // 11-th Conference on Empirical Methods in Natural Language Processing, 2011. — Pp. 227–237.

Повышение адекватности тематических моделей

Хорошая тематическая модель должна учитывать максимум дополнительной информации, особенности языка и коллекции:

- готовые классификаторы и рубрикаторы
- готовые словари терминов (ключевых фраз)
- связи документ — тема
- связи термин — тема
- фон — нетематические общеупотребительные слова
- шум — нетематические специфичные слова
- разреженность распределений $p(w|t)$ и $p(t|d)$
- явление усечения тем (burstiness)
- выделение n-грамм с учётом морфологии и синтаксиса

Требования к производительности

Реальные коллекции: $|D| \sim 10^{6...7}$, $|W| \sim 10^{4...6}$, $|T| \sim 10^{2...5}$.

- параллельные вычисления
- распределённое хранение
- онлайновая (однопроходная) обработка коллекции
- эффективные алгоритмы разреживания

Направления исследований научной группы ВЦ РАН

- Разработка *многофункциональной тематической модели* — иерархической, динамической, мультиязычной, робастной, разреженной, устойчивой, хорошо интерпретируемой, n-граммной, учитывающей особенности языка.
- *Big Data* — разработка технологии тематического моделирования больших коллекций документов.

Основные научные проблемы:

- Тематические модели на основе LDA и HDP слишком сложны для совмещения 3 и более требований.
- Построение качественной тематической иерархии.

Пути решения:

- Развитие более простой модели PLSA.
- Привлечение готовых рубрикаторов и словарей.

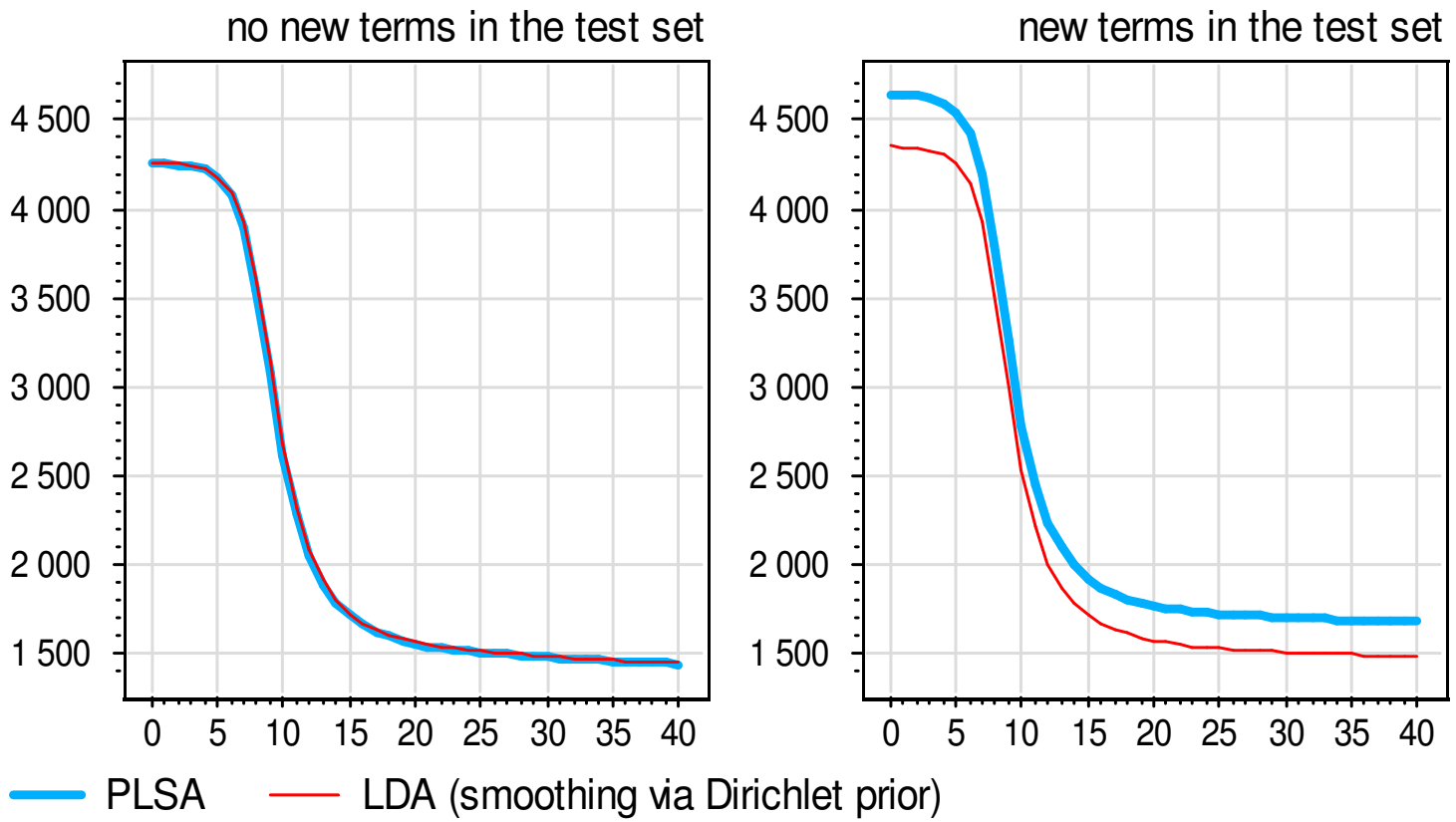
Текущее состояние исследований и результаты

На пути к многофункциональной модели:

- Построен обобщённый алгоритм оптимизации PLSA/LDA.
- Разработаны эффективные методы разреживания.
- Показано преимущество робастного разреженного PLSA.
- Построена онлайновая иерархическая модель на основе УДК и связей документ–тема.
- Строятся статистические тесты для анализа и коррекции структуры иерархии — выявления несоответствий документов, терминов, подтем их темам.

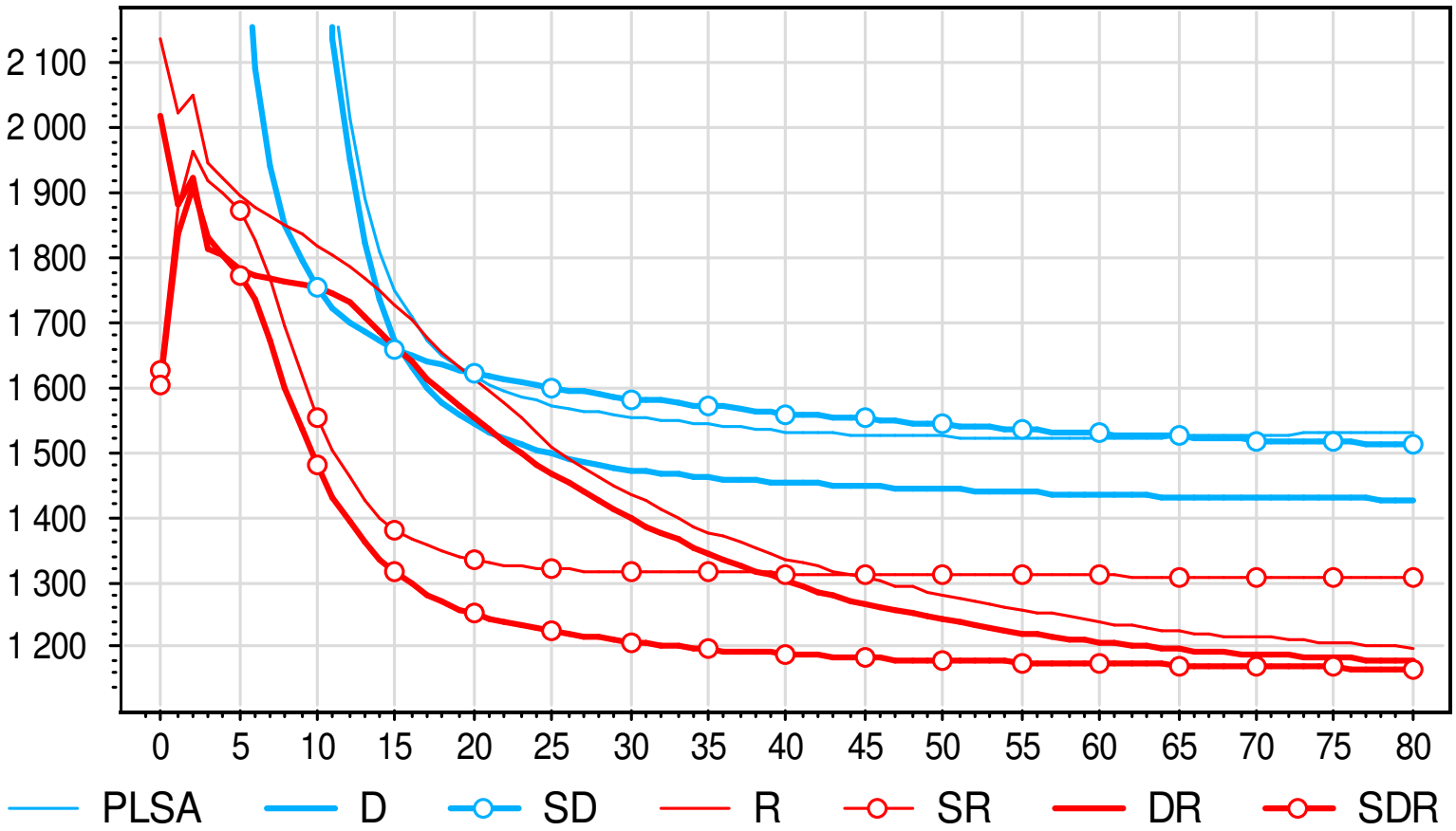
PLSA и LDA почти не отличаются по перплексии

Корректная перепроверка экспериментов (D. Blei 2003) показала, что LDA не имеет столь заметных преимуществ:



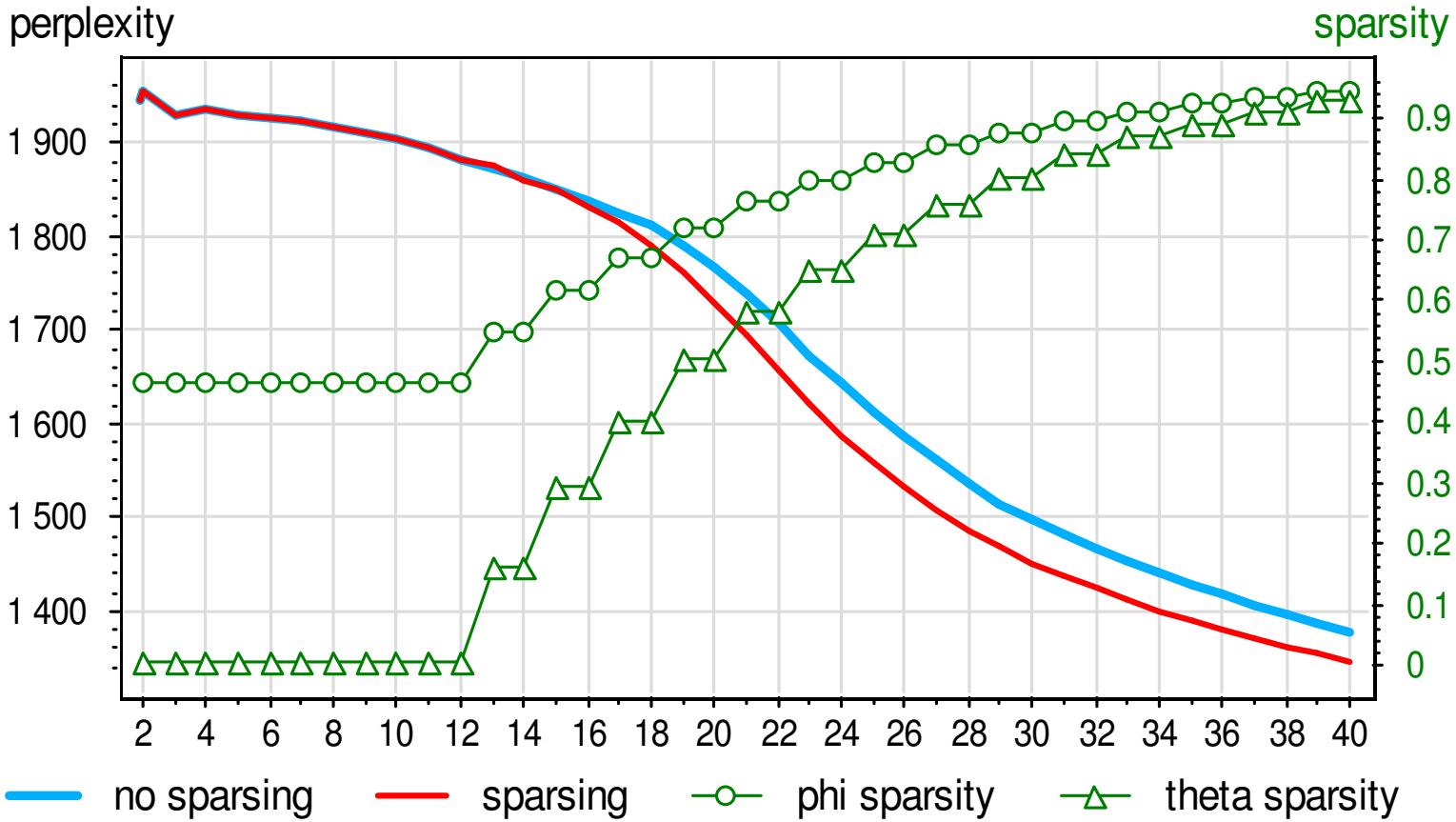
Робастность даёт заметный выигрыш в перплексии

Важен не только тип модели, но и метод её оптимизации (S — сэмплирование, D — Дирихле, R — робастность):



Сильное разреживание не ухудшает перплексию

Робастное разреживание $p(w|t)$, $p(t|d)$ — более 90% нулей.



Публикации

Potapenko A. A., Vorontsov K. V., Robust PLSA Performs Better Than LDA // European Conference on Information Retrieval ECIR-2013, Moscow, 24–27 March 2013. — Pp. 784–787.

Воронцов К. В., Потапенко А. А. Регуляризация, робастность и разреженность вероятностных тематических моделей // Компьютерные исследования и моделирование, 2012. — Т. 4, №12. — С. 693–706.

Царьков С. В. Автоматическое выделение ключевых фраз для построения словаря терминов в тематических моделях коллекций текстовых документов // Естественные и технические науки, 2012. — № 6. — С. 456–464.

Целых В. Р., Воронцов К. В. Критерии согласия для разреженных дискретных распределений и их применение в тематическом моделировании // Машинное обучение и анализ данных, 2012. — Т. 1, № 4. — С. 437–447.

Спасибо за внимание!

Воронцов Константин Вячеславович
voron@forecsys.ru

Страницы на www.MachineLearning.ru:

- Участник:Vokov
- Вероятностные тематические модели (курс лекций, К. В. Воронцов)
- Тематическое моделирование