• Вероятностные языковые модели • Лекция 10. Темпоральные, сегментирующие, транзакционные тематические модели

Константин Вячеславович Воронцов
k.vorontsov@iai.msu.ru

Этот курс доступен на странице вики-ресурса
http://www.MachineLearning.ru/wiki
«Вероятностные языковые модели (курс лекций, К.В.Воронцов)»

ВМК МГУ • 4 мая 2026

Содержание

Напоминание. Мультимодальная ARTM «мешка термов»

Дано: коллекция D, словари W_m модальностей $m \in M$
n_{dw} — частота терма $w \in W_m$ в документе $d \in D$.

Найти: вероятностную языковую модель $p(w|d) = \sum_{t \in T} \phi_{wt}\theta_{td}$
с параметрами $\phi_{wt} = p(w|t)$ и $\theta_{td} = p(t|d)$

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

EM-алгоритм: метод простой итерации для системы уравнений

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

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

Напоминание. Мультимодальная ARTM локальных контекстов

Дано: термы $w_1, \ldots, w_n$ и их контексты $C_i \subset \{1, \ldots, n\}$

Найти: вер. языковую модель $p(w|C_i) = \sum_{t \in T} \phi_{tw} \frac{p(w)}{p(t)} p(t|C_i)$

Критерий: $\sum_{i=1}^{n} \tau_{m_i} \ln \sum_{t \in T} \phi_{tw_i} \frac{p(w_i)}{p(t)} \sum_{c \in C_i} \alpha_{ci} \phi_{tw_c} + R(\Phi) \to \max_{\Phi}$

Контекст C_i зависит от модальности $m_i = m(w_i)$ терма w_i:

- слова, n-граммы, названия, ссылки — локальный контекст
- авторы, время, пользователи — весь документ, $\alpha_{ci} = \frac{1}{n_d}$

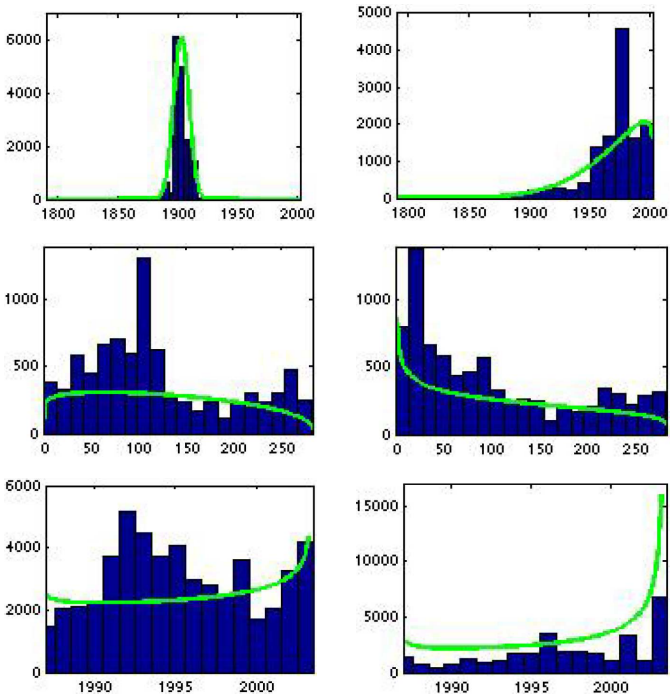
EM-алгоритм: метод простой итерации для системы уравнений

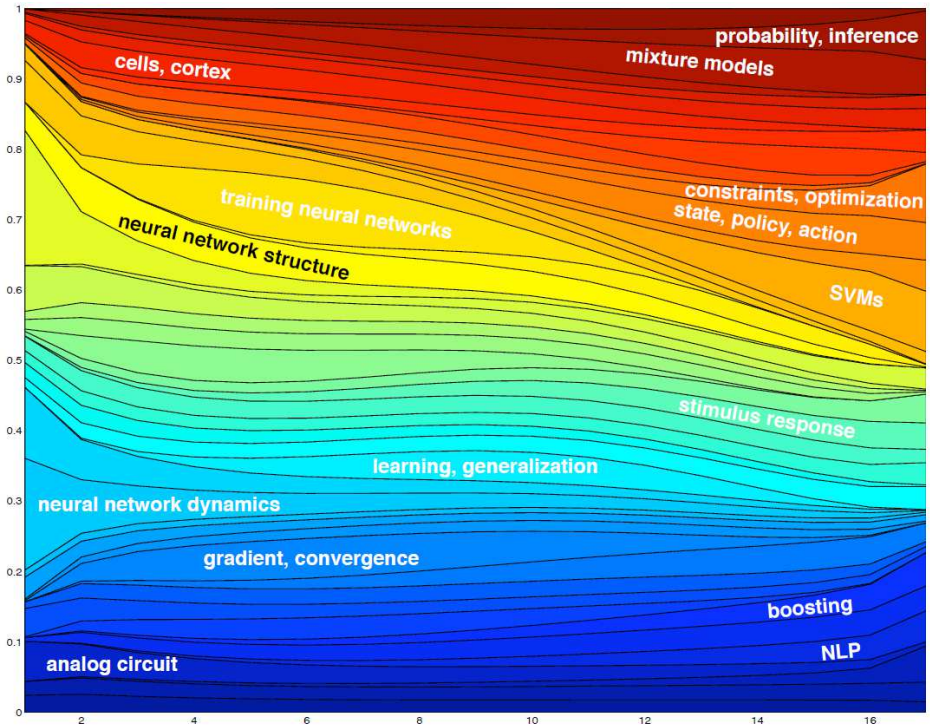
$$\text{E-шаг:} \quad p_{ti} = \operatorname*{norm}_{t \in T} \left(\frac{\phi_{tw_i}}{p(t)} \sum_{c \in C_i} \alpha_{ci} \phi_{tw_c} \right), \quad p(t) = \sum_{w \in W} \phi_{tw} p(w)$$

$$\text{M-шаг:} \quad \phi_{tw} = \operatorname*{norm}_{t \in T} \left(n_{tw} + \phi_{tw} \frac{\partial R}{\partial \phi_{tw}} \right), \quad n_{tw} = \sum_{i=1}^{n} \tau_{m_i} p_{ti} [w_i = w]$$

Модель TOT (Topics over Time)

1. Каждая тема имеет непрерывное β-распределение во времени
2. Каждое слово имеет метку времени





Xuerui Wang, Andrew McCallum. Topics over Time: A Non-Markov Continuous-Time Model of Topical Trends. ACM SIGKDD-2006

Темпоральные тематические модели

Неадекватность TOT очевидна даже по картинкам из статьи!

Наш подход. Предположения:

- Время дискретно, $i \in I$ — интервалы времени
- Как и в TOT, темы $p(w|t)$ не меняются во времени
- Метки времени приписываются документам, а не словам
- *Перманентные* темы имеют медленно меняющиеся $p(i|t)$
- *Событийные* темы имеют $p(i|t) = 0$ почти всё время
- Параметрические модели $p(i|t)$ не используются

Цели моделирования:

- Выделить событийные и перманентные темы
- Детектировать события (fist story / event detection)
- Проследить динамику развития тем во времени
- Выделить тренды (в новостях, в научных публикациях)

Регуляризаторы Θ для темпоральных тематических моделей

I — интервалы времени (например, годы публикаций),
$D_i \subset D$ — все документы, относящиеся к интервалу $i \in I$.
$n_i = \sum\limits_{d \in D_i} n_d$ — доля коллекции, относящаяся к интервалу i.

1. Разреживание $p(t|i) = \sum\limits_{d \in D_i} \theta_{td} \frac{n_d}{n_i}$ в каждом интервале i:

$$R_{\text{разр}}(\Theta) = \tau_{\text{разр}} \sum_{i \in I} \text{KL}\left(\tfrac{1}{|T|} \| p(t|i)\right) \to \max.$$

2. Сглаживание $p(i|t) = \sum\limits_{d \in D_i} \theta_{td} \frac{n_d}{n_t}$ в соседних интервалах i, $i-1$:

$$R_{\text{сгл}}(\Theta) = -\tau_{\text{сгл}} \sum_{i \in I} \sum_{t \in T} \left| p(i|t) - p(i-1|t) \right| \to \max.$$

Seung-Jean Kim, Kwangmoo Koh, Stephen Boyd, Dimitry Gorinevsky. L1 trend filtering. SIAM review, 2009.

Время как модальность. Регуляризаторы Ф

Проблема регуляризатора Θ в пакетном EM-алгоритме: соседние по времени документы могут попасть в разные пакеты.

Документы содержат слова $w \in W^1$ и время $i \in W^2 = I$
W^2 — модальность интервалов времени (time stamps)

1. Разреживание $p(t|i)$ эквивалентно разреживанию $p(i|t) = \phi_{it}$:

$$R_{\text{разр}}(\Phi_2) = -\tau_{\text{разр}} \sum_{i \in I} \sum_{t \in T} \ln \phi_{it} \to \max$$

2. Сглаживание $p(i|t) = \phi_{it}$ в соседних интервалах i, $i-1$:

$$R_{\text{сгл}}(\Phi_2) = -\tau_{\text{сгл}} \sum_{i \in I} \sum_{t \in T} \left|\phi_{it} - \phi_{i-1,t}\right| \to \max$$

Мультимодальная ARTM с суммой L_1-регуляризаторов

Максимизация суммы log-правдоподобий с регуляризацией:

$$\sum_{m,d,w} \tau_m n_{dw} \ln \sum_{t \in T} \phi_{wt}\theta_{td} + R(\Phi,\Theta) - \sum_{j \in J} \lambda_j \left| R_j(\Phi,\Theta) \right| \to \max_{\Phi,\Theta}$$

EM-алгоритм: метод простой итерации для системы уравнений

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

M-шаг:
$$\begin{cases} \phi_{wt} = \underset{w \in W^m}{\text{norm}}\left(n_{wt} + \phi_{wt}\frac{\partial R}{\partial \phi_{wt}} - \phi_{wt}\sum_{j \in J} \lambda_j \operatorname{sign}(R_j) \frac{\partial R_j}{\partial \phi_{wt}} \right) \\ \theta_{td} = \underset{t \in T}{\text{norm}}\left(n_{td} + \theta_{td}\frac{\partial R}{\partial \theta_{td}} - \theta_{td}\sum_{j \in J} \lambda_j \operatorname{sign}(R_j) \frac{\partial R_j}{\partial \theta_{td}} \right); \end{cases}$$

Никита Дойков. Адаптивная регуляризация вероятностных тематических моделей // ВКР бакалавра, 2015. ВМК МГУ.

Сглаживающий L_1-регуляризатор временного ряда

подходит для интерполяции разрывных временных рядов, т.к.

- не штрафует модель за резкие скачки,
- не сглаживает $p(i|t)$ в момент i появления новой темы,
- в отличие от сглаживающего L_2-регуляризатора

Формула M-шага для модальности времени с $R_{\text{сгл}}(\Phi)$:

$$\phi_{it} = \underset{i \in I}{\text{norm}}\Big(n_{it} - \tau_{\text{сгл}}\phi_{it}\big(\text{sign}(\phi_{it} - \phi_{i-1,t}) + \text{sign}(\phi_{it} - \phi_{i+1,t})\big)\Big)$$

- если ϕ_{it} выше соседних $\phi_{i\pm1,t}$, то ϕ_{it} уменьшается
- если ϕ_{it} ниже соседних $\phi_{i\pm1,t}$, то ϕ_{it} увеличивается
- если ϕ_{it} попадает между ними, то ϕ_{it} не изменяется

Seung-Jean Kim, Kwangmoo Koh, Stephen Boyd, Dimitry Gorinevsky.
L1 trend filtering. SIAM review, 2009.

Задача анализа потока пресс-релизов

Коллекция официальных пресс-релизов внешнеполитических ведомств ряда стран на английском языке.
Более 20 тыс. сообщений за 10 лет, 180Мб текста.

Цели исследования:

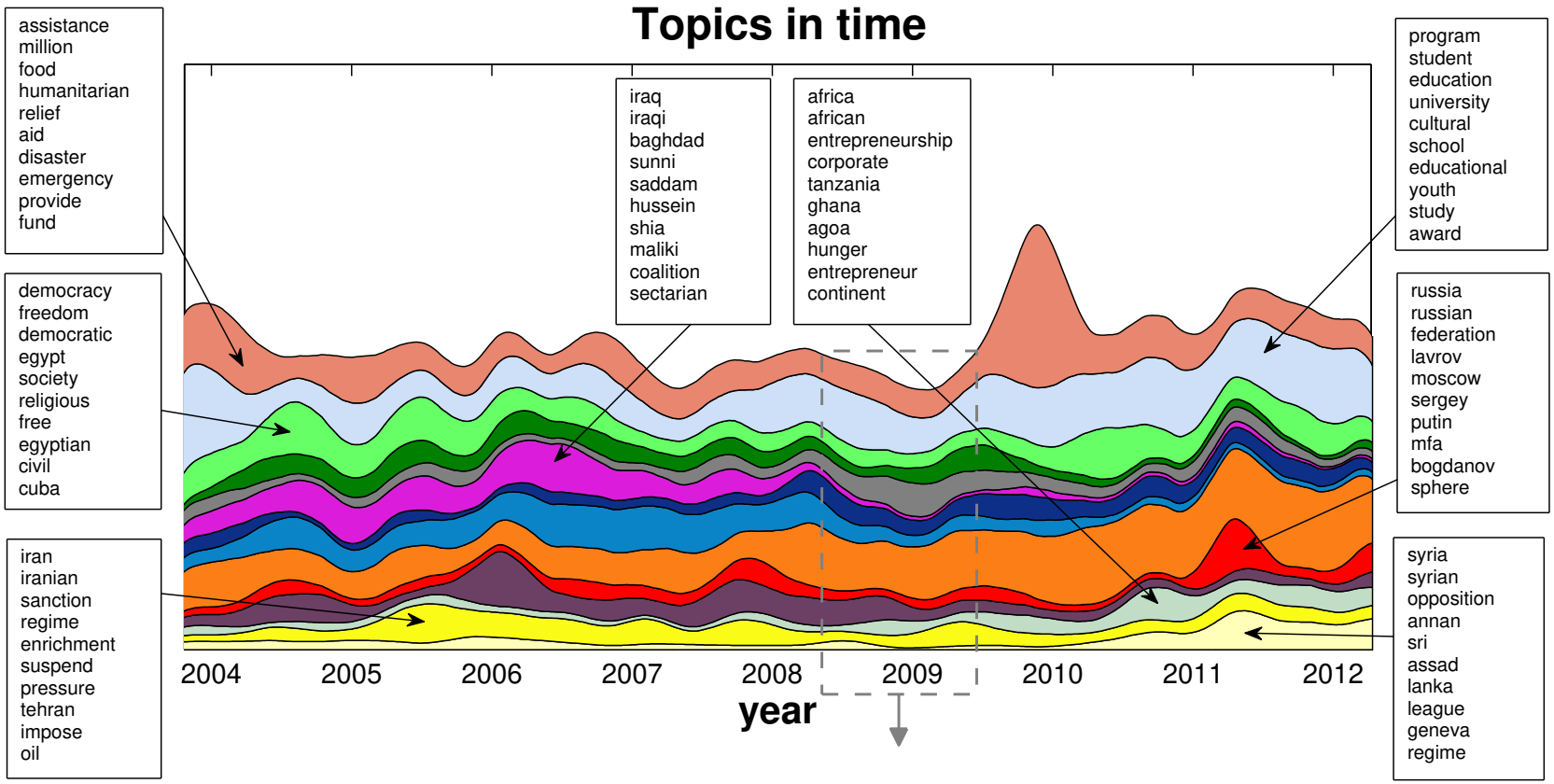
- какие темы общие, какие специфичны для источников?
- какие темы событийные, какие перманентные?
- какие темы и когда коррелируют с заданной темой?

Модальности и регуляризаторы:

- две модальности: источники, интервалы времени
- разреживание, сглаживание, декоррелирование
- сглаживание тем во времени

Никита Дойков. Адаптивная регуляризация вероятностных тематических моделей // ВКР бакалавра, 2015. ВМК МГУ.

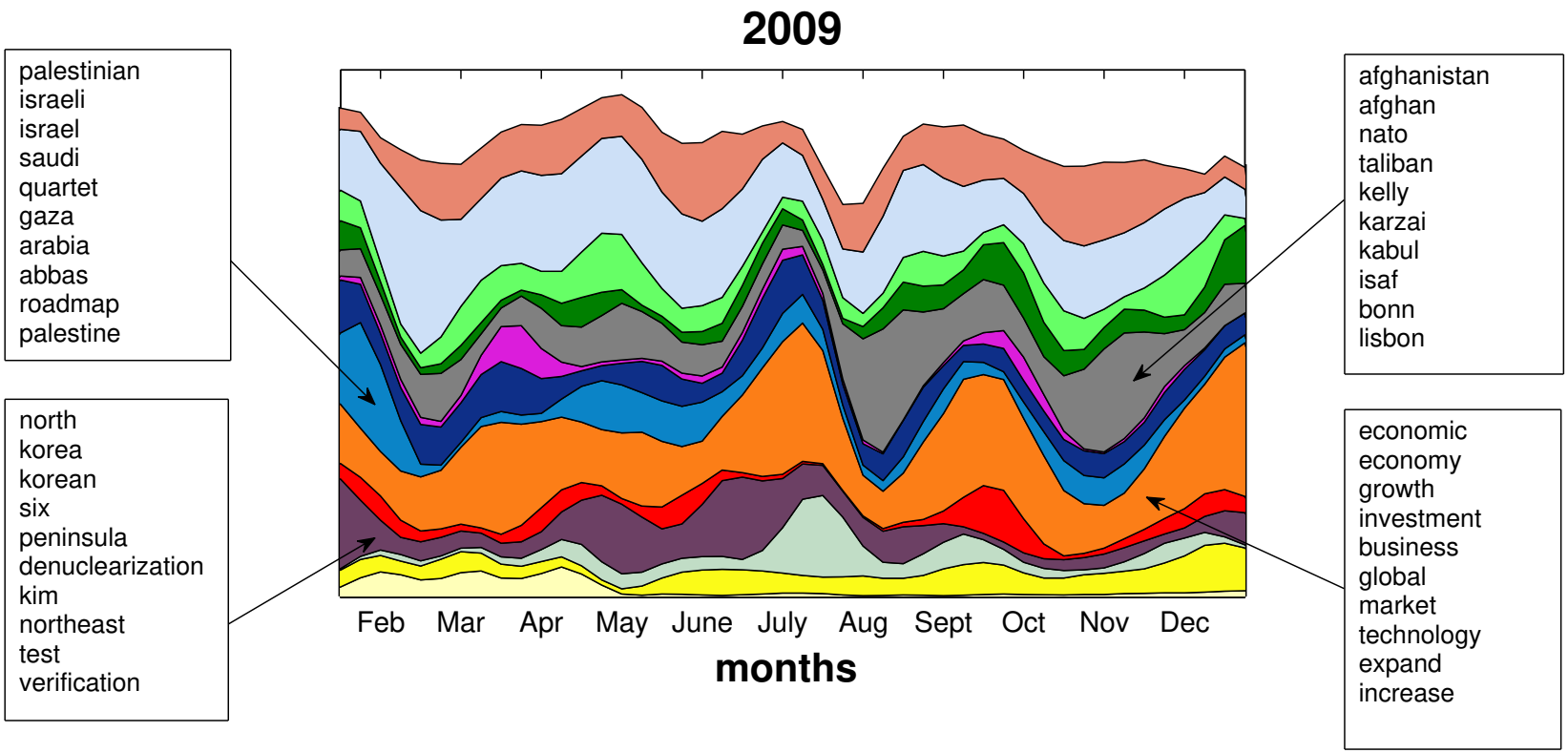
Динамика тем во времени



Никита Дойков. Адаптивная регуляризация вероятностных тематических моделей // ВКР бакалавра, 2015. ВМК МГУ.

Динамика тем во времени

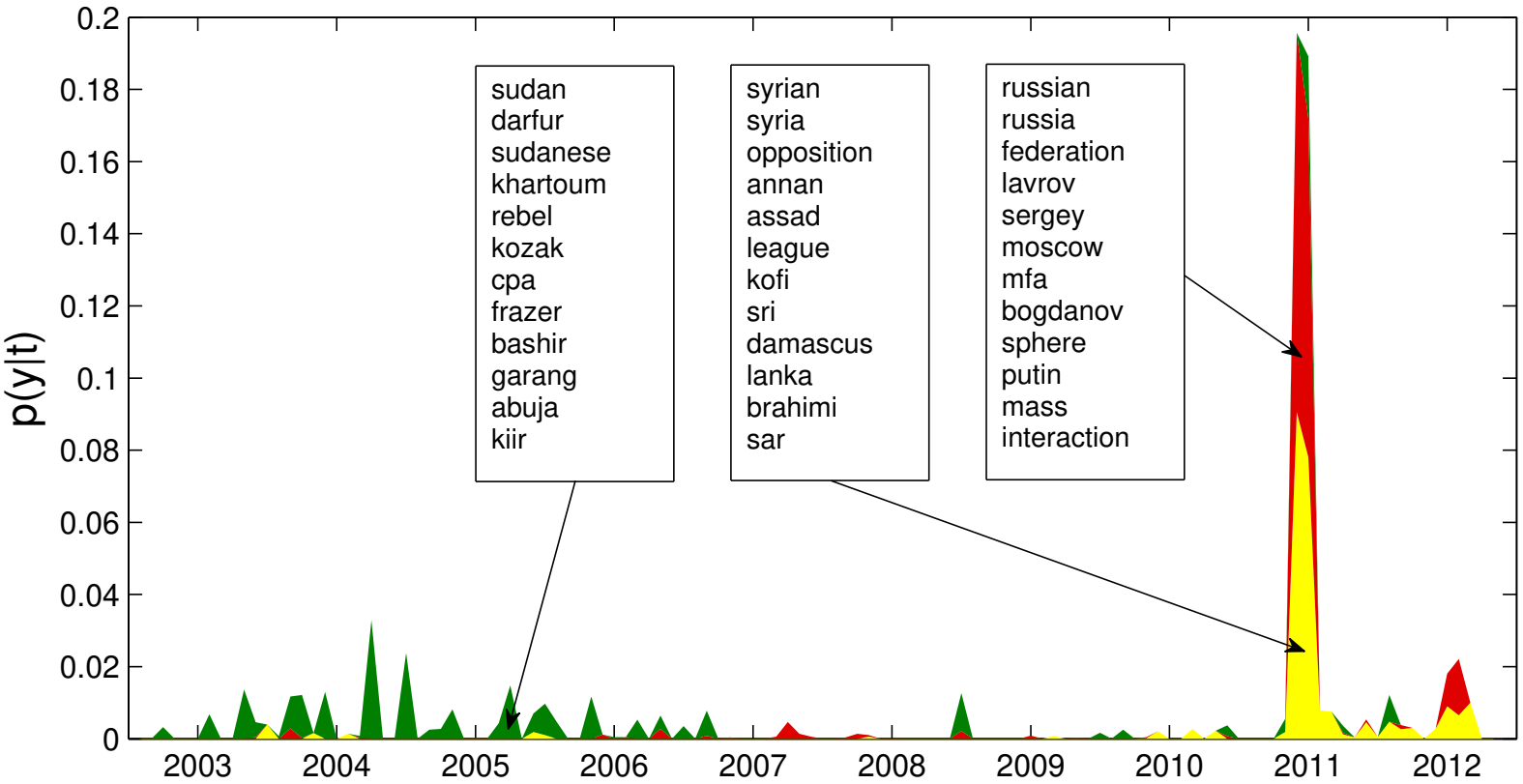
Укрупнение масштаба времени



Никита Дойков. Адаптивная регуляризация вероятностных тематических моделей // ВКР бакалавра, 2015. ВМК МГУ.

Динамика тем во времени

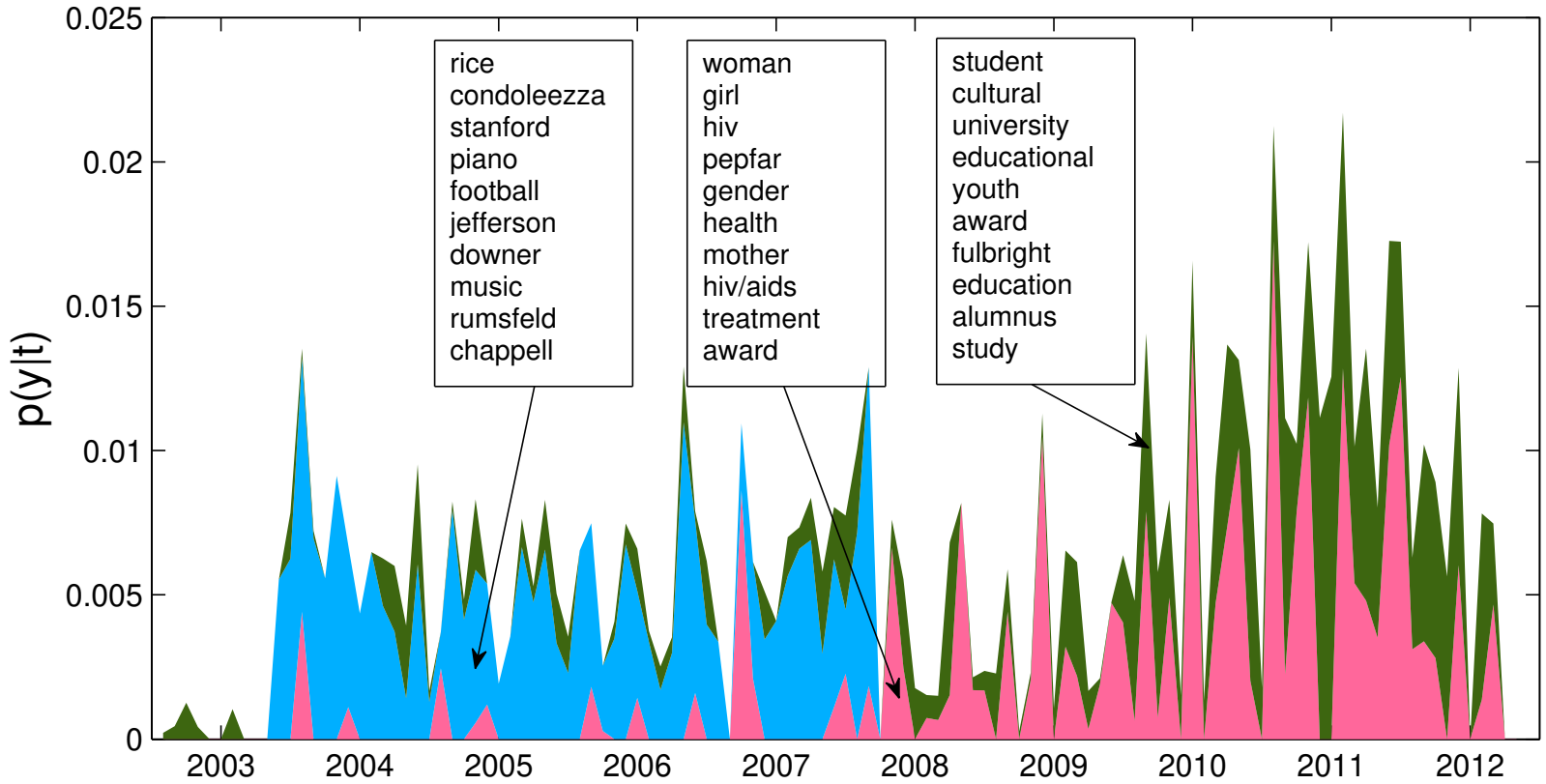
Пример: событийные темы и момент их совместного всплеска



Никита Дойков. Адаптивная регуляризация вероятностных тематических моделей // ВКР бакалавра, 2015. ВМК МГУ.

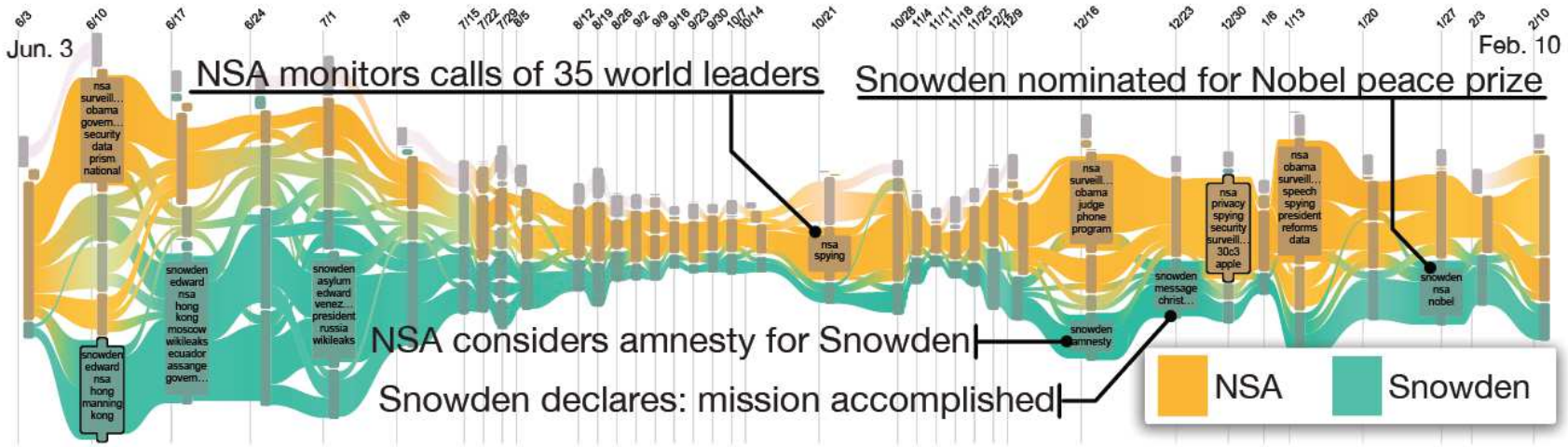
Динамика тем во времени

Примеры перманентных тем (сглаживание отключено)



Никита Дойков. Адаптивная регуляризация вероятностных тематических моделей // ВКР бакалавра, 2015. ВМК МГУ.

Динамика тем: эволюция предметной области



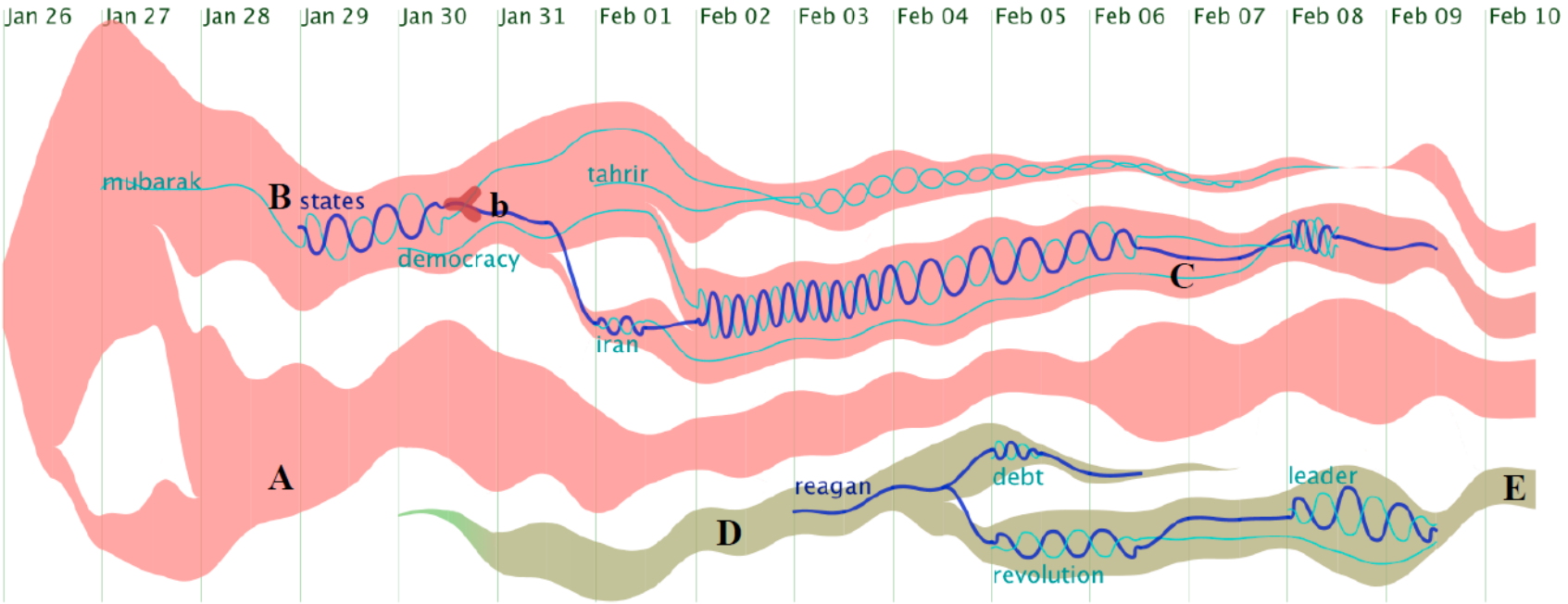
Эволюция выбранных тем иерархии. Данные Prism (2013/06/03–2014/02/09)

Стратегия визуализации в системах TextFlow и RoseRiver:

- эксперт задаёт сечение иерархии (дерева) тем,
- интерактивно выбирает подмножество тем и событий,
- получает сгенерированный отчёт с инфографикой.

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

Пример динамической модели

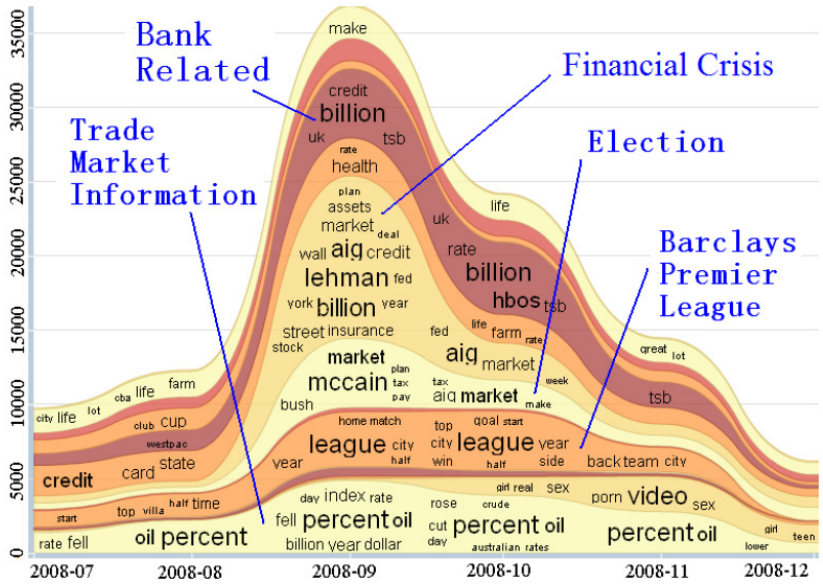


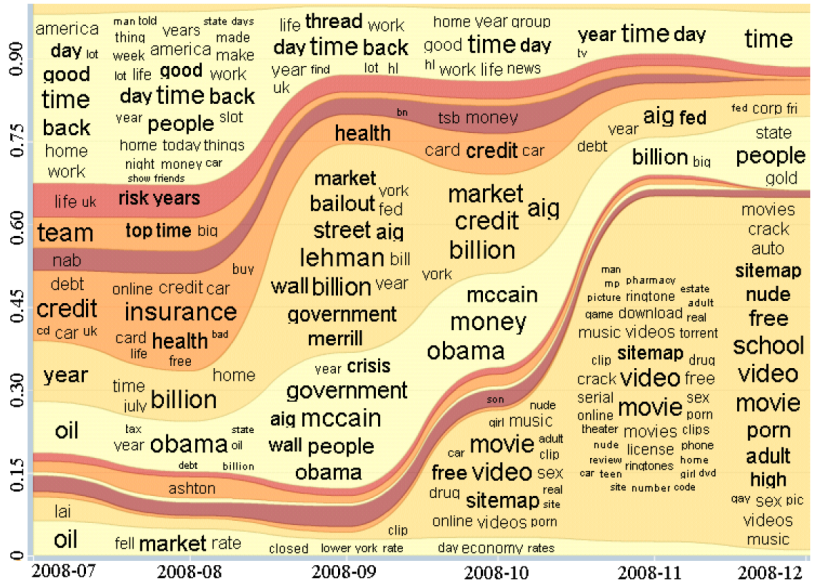
Выявляются и отображаются:

- моменты разделения и слияния тем
- критические события; подтемы или нити повествования
- корреляции между частотами ключевых слов

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. 2011.

Ещё пример динамической модели





Выявляются и отображаются:

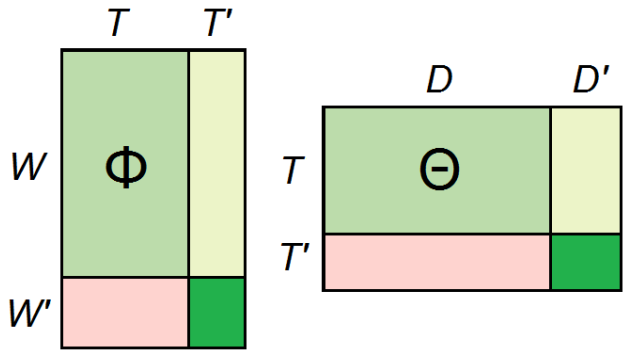
- динамика тем по новостным источникам
- «облака слов» по их значимости в динамике

Jianwen Zhang, Yangqiu Song, Changshui Zhang, Shixia Liu. Evolutionary hierarchical Dirichlet processes for multiple correlated time-varying corpora. 2010.

Онлайновые темпоральные модели

При добавлении нового пакета D' в коллекцию D добавляются:

- новые слова W' в W
- новые темы T' в T



Как обнаружить в $d \in D'$ новую тему? (First Story Detection)
Как понять, что это продолжение? (Topic Detection & Tracking)
Как понять, сколько различных новых тем есть в D'?
Как обеспечить полноту и точность финального набора тем?
Как своевременно обнаружить *тренд* — растущую тему?

James Allan et al. Topic detection and tracking pilot study: Final report. DARPA, 1998

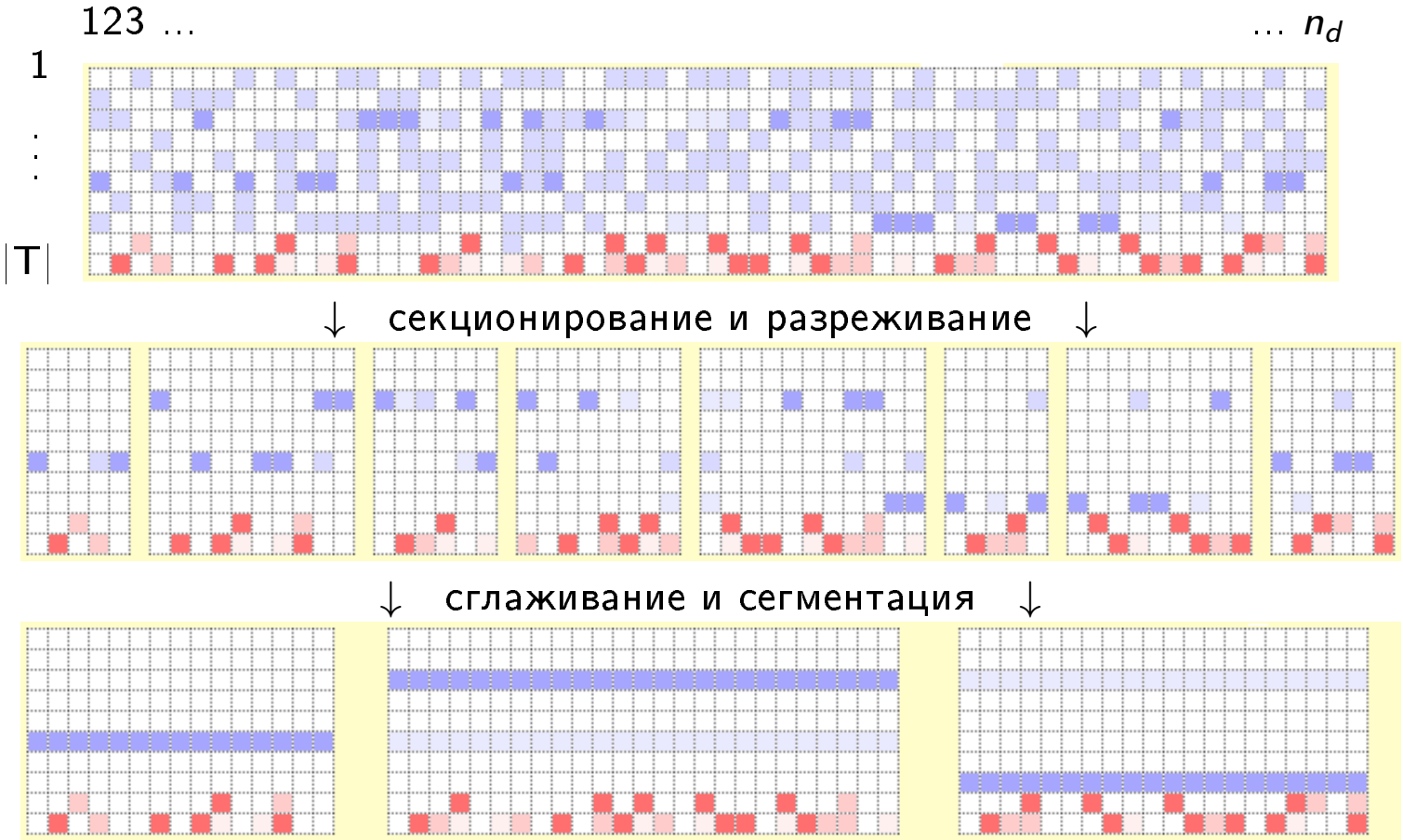
James Allan. Topic detection and tracking: Event-based information organization. 2002

Н.Герасименко, А.Чернявский, М.Никифорова, М.Никитин, К.Воронцов. Инкрементальное обучение тематических моделей для поиска трендовых тем в научных публикациях. Доклады РАН, 2022.

Сегментная структура текста и пост-обработка E-шага

Документ $d = \{w_1, \ldots, w_{n_d}\}$, n_d — длина документа d

Тематика термов в документе $p(t|d, w_i)$ — матрица $T \times n_d$:



Регуляризация E-шага

Трёхмерная матрица $\Pi = \left(p_{tdw} = p(t|d,w)\right)_{T\times D\times W}$

Регуляризатор E-шага: $\tilde{R}(\Phi,\Theta) = R(\Pi(\Phi,\Theta),\Phi,\Theta)$:

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

EM-алгоритм: метод простой итерации для системы уравнений

$$\begin{cases} \text{E-шаг:} & p_{tdw} = \underset{t\in T}{\text{norm}}\left(\phi_{wt}\theta_{td}\right) \\ & \tilde{p}_{tdw} = p_{tdw}\left(1 + \frac{1}{n_{dw}}\left(\frac{\partial R}{\partial p_{tdw}} - \sum_{z\in T} p_{zdw}\frac{\partial R}{\partial p_{zdw}}\right)\right) \qquad (*) \\ \text{M-шаг:} & \phi_{wt} = \underset{w\in W}{\text{norm}}\left(\sum_{d\in D} n_{dw}\tilde{p}_{tdw} + \phi_{wt}\frac{\partial R}{\partial \phi_{wt}}\right) \\ & \theta_{td} = \underset{t\in T}{\text{norm}}\left(\sum_{w\in d} n_{dw}\tilde{p}_{tdw} + \theta_{td}\frac{\partial R}{\partial \theta_{td}}\right) \end{cases}$$

Набросок доказательства: три шага

1. Для функции $p_{tdw}(\Phi,\Theta) = \frac{\phi_{wt}\theta_{td}}{\sum_z \phi_{wz}\theta_{zd}}$ и любого $z \in T$

$$\phi_{wt}\frac{\partial p_{zdw}}{\partial \phi_{wt}} = \theta_{td}\frac{\partial p_{zdw}}{\partial \theta_{td}} = p_{tdw}\big([z=t] - p_{zdw}\big).$$

2. Введём вспомогательную функцию от переменных Π, Φ, Θ:

$$Q_{tdw}(\Pi,\Phi,\Theta) = \frac{\partial R(\Pi,\Phi,\Theta)}{\partial p_{tdw}} - \sum_{z\in T} p_{zdw}\frac{\partial R(\Pi,\Phi,\Theta)}{\partial p_{zdw}}.$$

Если $R(\Pi,\Phi,\Theta)$ не зависит от p_{tdw} при $w \notin d$, то

$$\phi_{wt}\frac{\partial \tilde{R}}{\partial \phi_{wt}} = \phi_{wt}\frac{\partial R}{\partial \phi_{wt}} + \sum_{d\in D} p_{tdw}Q_{tdw};\quad \theta_{td}\frac{\partial \tilde{R}}{\partial \theta_{td}} = \theta_{td}\frac{\partial R}{\partial \theta_{td}} + \sum_{w\in d} p_{tdw}Q_{tdw}.$$

3. Подставляем это в формулы M-шага:

$$\phi_{wt} = \underset{w\in W}{\text{norm}}\Big(\sum_{d\in D} n_{dw}p_{tdw} + \sum_{d\in D} Q_{tdw}p_{tdw} + \phi_{wt}\frac{\partial R}{\partial \phi_{wt}}\Big);$$

$$\theta_{td} = \underset{t\in T}{\text{norm}}\Big(\sum_{w\in d} n_{dw}p_{tdw} + \sum_{w\in d} Q_{tdw}p_{tdw} + \theta_{td}\frac{\partial R}{\partial \theta_{td}}\Big).$$

Шаг 1. Замечательное тождество

Для функции $p_{tdw}(\Phi,\Theta) = \frac{\phi_{wt}\theta_{td}}{\sum_z \phi_{wz}\theta_{zd}}$ и любого $z \in T$

$$\phi_{wt}\frac{\partial p_{zdw}}{\partial \phi_{wt}} = \theta_{td}\frac{\partial p_{zdw}}{\partial \theta_{td}} = p_{tdw}\big([z=t] - p_{zdw}\big).$$

Воспользуемся определением функции $p_{tdw}(\Phi,\Theta)$:

$$\phi_{wt}\frac{\partial p_{zdw}}{\partial \phi_{wt}} = \phi_{wt}\frac{[z=t]\theta_{td}\sum_u \phi_{wu}\theta_{ud} - \theta_{td}\phi_{wz}\theta_{zd}}{(\sum_u \phi_{wu}\theta_{ud})^2} =$$

$$= p_{tdw}[z=t] - p_{tdw}p_{zdw} = p_{tdw}\big([z=t] - p_{zdw}\big);$$

$$\theta_{td}\frac{\partial p_{zdw}}{\partial \phi_{td}} = \theta_{td}\frac{[z=t]\phi_{wt}\sum_u \phi_{wu}\theta_{ud} - \phi_{wt}\phi_{wz}\theta_{zd}}{(\sum_u \phi_{wu}\theta_{ud})^2} =$$

$$= p_{tdw}[z=t] - p_{tdw}p_{zdw} = p_{tdw}\big([z=t] - p_{zdw}\big).$$

Шаг 2. Дифференцирование суперпозиции $R(\Pi(\Phi,\Theta),\Phi,\Theta)$

Пусть $R(\Pi)$ не зависит от переменных p_{tdw} при $w \notin d$. Тогда

$$\phi_{wt}\frac{\partial \tilde{R}}{\partial \phi_{wt}} = \phi_{wt}\frac{\partial R}{\partial \phi_{wt}} + \sum_d p_{tdw} Q_{tdw};$$

$$\theta_{td}\frac{\partial \tilde{R}}{\partial \theta_{td}} = \theta_{td}\frac{\partial R}{\partial \theta_{td}} + \sum_w p_{tdw} Q_{tdw}; \quad Q_{tdw} = \frac{\partial R}{\partial p_{tdw}} - \sum_{z\in T} p_{zdw}\frac{\partial R}{\partial p_{zdw}}$$

Заметим: $\frac{\partial p_{zdw'}}{\partial \phi_{wt}} = 0,\ w \neq w';\quad \frac{\partial p_{zd'w}}{\partial \theta_{td}} = 0,\ d \neq d';\quad \frac{\partial R}{\partial p_{tdw}} = 0,\ w \notin d.$

$$\phi_{wt}\frac{\partial \tilde{R}}{\partial \phi_{wt}} = \phi_{wt}\left(\frac{\partial R}{\partial \phi_{wt}} + \sum_{z,d,w'} \frac{\partial R}{\partial p_{zdw'}}\frac{\partial p_{zdw'}}{\partial \phi_{wt}}\right) = \phi_{wt}\frac{\partial R}{\partial \phi_{wt}} + \sum_{d,z} \frac{\partial R}{\partial p_{zdw}}\phi_{wt}\frac{\partial p_{zdw}}{\partial \phi_{wt}}$$

$$\theta_{td}\frac{\partial \tilde{R}}{\partial \theta_{td}} = \theta_{td}\left(\frac{\partial R}{\partial \theta_{td}} + \sum_{z,d',w} \frac{\partial R}{\partial p_{zd'w}}\frac{\partial p_{zd'w}}{\partial \theta_{td}}\right) = \theta_{td}\frac{\partial R}{\partial \theta_{td}} + \sum_{w,z} \frac{\partial R}{\partial p_{zdw}}\theta_{td}\frac{\partial p_{zdw}}{\partial \theta_{td}}$$

В силу «замечательного тождества» шага 1

$$\sum_{z\in T} \frac{\partial R}{\partial p_{zdw}} p_{tdw}\big([z=t] - p_{zdw}\big) = p_{tdw} Q_{tdw}.$$

Шаг 3. Подстановка производных $\tilde{R}(\Phi, \Theta)$ в формулы M-шага

Точка максимума (Φ, Θ) регуляризованного log-правдоподобия

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

удовлетворяет системе уравнений относительно ϕ_{wt}, θ_{td}, p_{tdw}:

$$p_{tdw} = \operatorname*{norm}_{t \in T} \left(\phi_{wt} \theta_{td} \right);$$

$$\phi_{wt} = \operatorname*{norm}_{w \in W} \left(\sum_{d \in D} n_{dw} p_{tdw} + \sum_{d \in D} Q_{tdw} p_{tdw} + \phi_{wt} \frac{\partial R}{\partial \phi_{wt}} \right);$$

$$\theta_{td} = \operatorname*{norm}_{t \in T} \left(\sum_{w \in d} n_{dw} p_{tdw} + \sum_{w \in d} Q_{tdw} p_{tdw} + \theta_{td} \frac{\partial R}{\partial \theta_{td}} \right).$$

Общий член в формулах M-шага переносится в E-шаг, если ввести новую переменную $\tilde{p}_{tdw} = p_{tdw} \left(1 + \frac{1}{n_{dw}} Q_{tdw} \right)$. ■

Любая пост-обработка E-шага — это регуляризатор $R(\Pi)$

Итак, произвольному гладкому регуляризатору $R(\Pi, \Phi, \Theta)$ однозначно соответствует пост-обработка $p_{tdw} \to \tilde{p}_{tdw}$.

Оказывается, верно и обратное:

Теорема. Если на k-й итерации EM-алгоритма для каждого (d, w): $n_{dw} > 0$ в формулах M-шага вместо вектора $(p^k_{tdw})_{t \in T}$ подставить вектор $(\tilde{p}^k_{tdw})_{t \in T}$, удовлетворяющий условию нормировки $\sum_t \tilde{p}^k_{tdw} = 1$, то это эквивалентно добавлению регуляризатора сглаживания–разреживания

$$R(\Pi) = \sum_{d \in D} \sum_{w \in d} n_{dw} \sum_{t \in T} (\tilde{p}^k_{tdw} - p^k_{tdw}) \ln p_{tdw}.$$

$p(t|d, w)$ можно подвергать любой разумной пост-обработке!
ОГО! И ТАК МОЖНО БЫЛО?!

Доказательство

В системе $(*)$ дифф. уравнений относительно R введём переменные x_{tdw}:

$$\underbrace{p_{tdw}^k \frac{\partial R}{\partial p_{tdw}}}_{x_{tdw}} = n_{dw}(\tilde{p}_{tdw}^k - p_{tdw}^k) + p_{tdw}^k \sum_{z \in T} \underbrace{p_{zdw}^k \frac{\partial R}{\partial p_{zdw}}}_{x_{zdw}}, \quad t \in T.$$

Для любой пары (d, w) такой, что $n_{dw} > 0$, это система $|T|$ линейных уравнений относительно $|T|$ переменных x_{tdw}, $t \in T$.
Подстановкой убеждаемся, что $x_{tdw} = n_{dw}(\tilde{p}_{tdw}^k - p_{tdw}^k)$ — решение системы.
Взяв это решение, получим систему дифф. уравнений относительно R:

$$\frac{\partial R}{\partial p_{tdw}} = \frac{x_{tdw}}{p_{tdw}}, \quad d \in D, \ w \in d, \ t \in T.$$

Система декомпозируется по переменным p_{tdw}: каждой тройке (d, w, t) соответствует частное решение $R(\Pi) = x_{tdw} \ln p_{tdw} + C$. Общее решение:

$$R(\Pi) = \sum_{d \in D} \sum_{w \in d} \sum_{t \in T} x_{tdw} \ln p_{tdw} + C.$$

Подставляя сюда найденное решение x_{tdw}, получаем требуемое. ■

Пример 1. Кросс-энтропийное разреживание $p(t|d,w)$

Путь каждый терм относится к небольшому числу тем:

$$\mathrm{KL}\Big(\tfrac{1}{|T|} \,\Big\|\, p(t|d,w)\Big) \to \max.$$

Суммируем по всем термам всех документов:

$$R(\Pi) = -\frac{\tau}{|T|}\sum_{d\in D}\sum_{w\in d} n_{dw}\sum_{t\in T}\ln p_{tdw} \to \max.$$

Формула регуляризованного E-шага:

$$\tilde{p}_{tdw} = p_{tdw} - \tau\Big(\frac{1}{|T|} - p_{tdw}\Big).$$

Интерпретация: Если $p_{tdw} < \frac{1}{|T|}$, то p_{tdw} станет ещё меньше.
Тематика терма концентрируется в небольшом числе тем.

Недостаток: Соседние векторы разреживаются независимо.

Пример 2. Тематическая модель сегментированного текста

S_d — множество сегментов (предложений) документа d
n_{sw} — число вхождений терма w в сегмент s длины n_s
Тематика сегмента $s \in S_d$ — среднее по всем его термам:

$$p_{tds} \equiv p(t|d,s) = \frac{1}{n_s} \sum_{w \in s} n_{sw} p_{tdw}.$$

Кросс-энтропийный регуляризатор разреживания $p(t|d,s)$:

$$R(\Pi) = -\sum_{d \in D} \sum_{s \in S_d} \sum_{t \in T} \ln \sum_{w \in s} n_{sw} p_{tdw} \to \max.$$

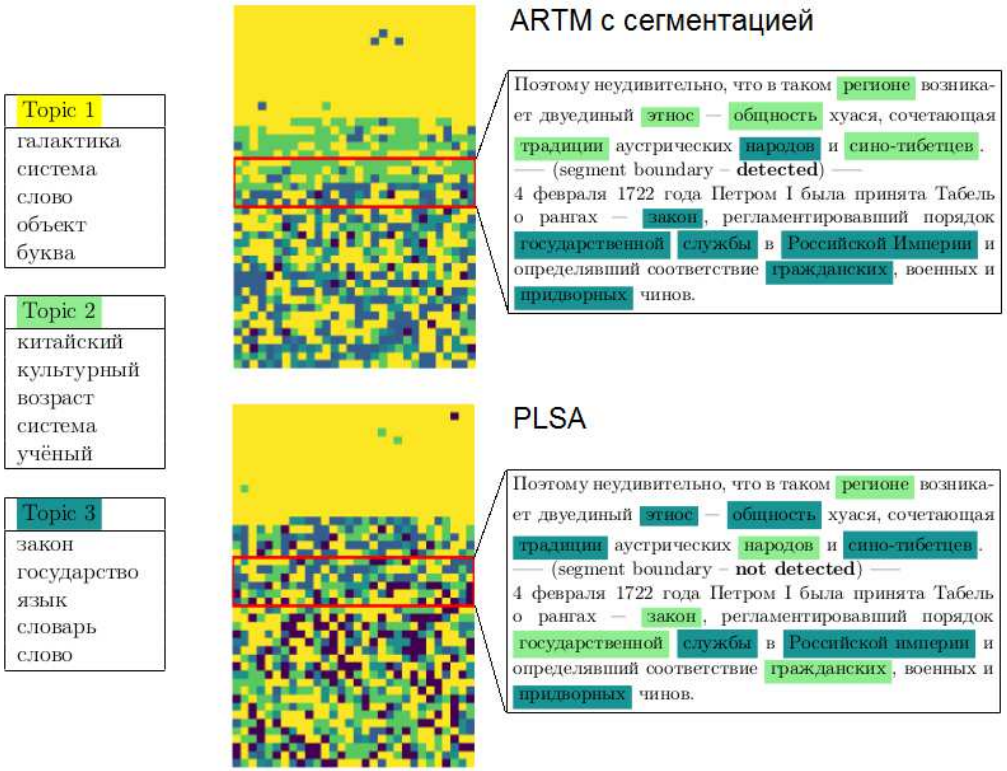
Формула регуляризованного E-шага:

$$\tilde{p}_{tdw} = p_{tdw} \left(1 - \frac{\tau}{n_{dw}} \sum_{s \in S_d} \frac{n_{sw}}{n_s} \left(\frac{1}{p_{tds}} - \sum_{z \in T} \frac{p_{zdw}}{p_{zds}}\right)\right).$$

Интерпретация: если $p_{tds} < \frac{1}{|T|}$, то p_{tdw} уменьшатся $\forall\, w \in s$.
Тематика сегмента концентрируется в небольшом числе тем.

Пример 2. Эксперимент на полусинтетической коллекции

Сегментация текстов, склеенных из сегментов монотематичных статей научно-просветительского портала postnauka.ru



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

Транзакционные данные

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

- **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
- **Данные о пассажирских авиаперелётах:**
  (u, a, b, c) — перелёт клиента u из a в b авиакомпанией c

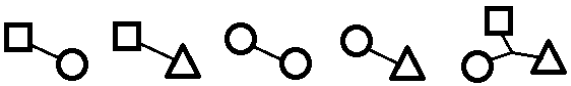
Задача: по наблюдаемой выборке рёбер гиперграфа найти латентные тематические векторные представления его вершин.

Тематическая модель гиперграфа: определения и обозначения

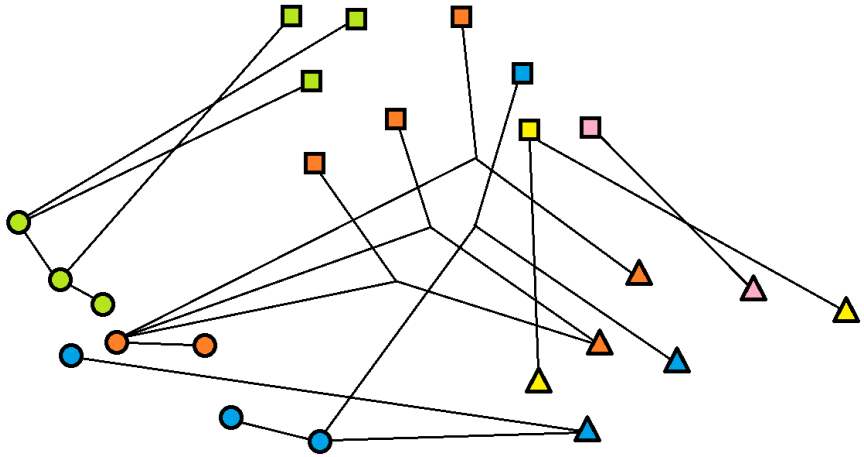
$\Gamma = \langle V, E \rangle$ — ориентированный гиперграф.
$V = V^1 \sqcup \cdots \sqcup V^M$ — разбиение вершин по модальностям

M — множество модальностей:

K — множество типов рёбер:



T — множество тем:



E_k — исходные данные: выборка рёбер-транзакций типа k
(d, x) — ребро: вершина-контейнер $d \in V$ и вершины $x \subset V$
n_{kdx} — число вхождений ребра (d, x) в выборку E_k

K.V.Vorontsov. Rethinking probabilistic topic modeling from the point of view of classical non-Bayesian regularization // Data Analysis and Optimization, Springer, 2023.

Тематическая модель гиперграфа: основные предположения

- в ребре (d, x) подмножество $x \subset V$ может быть любым, независимо от типа ребра k
- первая *гипотеза условной независимости*: тематика контейнера $p(t|d)$ не зависит от типа ребра k
- вторая *гипотеза условной независимости*: распределение $p(v|t)$ термов v модальности V^m в теме t не зависит ни от контейнера d, ни от типа ребра k
- третья *гипотеза условной независимости*: термы $v \in x$ в ребре (d, x) не зависят друг от друга
- *гипотеза «мешка транзакций»*: выборка транзакций типа k порождается случайно и независимо из

$$p_k(d, x) = p(d) \sum_{t \in T} p(t|d) \prod_{v \in x} p(v|t)$$

Тематическая модель гиперграфа

Вероятностная тематическая модель рёбер типа k

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

Задача максимизации взвешенной суммы log-правдоподобий по всем типам рёбер:

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

при ограничениях неотрицательности и нормировки:

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

где $\tau_k > 0$ — веса типов рёбер.

EM-алгоритм для гиперграфовой ARTM

Задача максимизации регуляризованного правдоподобия:

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

EM-алгоритм: метод простой итерации для системы уравнений со вспомогательными переменными $p_{tdx} = p(t|d,x)$:

$$\text{E-шаг:}\quad p_{tdx} = \underset{t\in T}{\text{norm}}\Big(\theta_{td}\prod_{v\in x}\phi_{vt}\Big)$$

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

Доказательство (по лемме о максимизации на симплексах)

Применим Лемму к log-правдоподобию с регуляризатором R:

$$\phi_{vt} = \underset{v \in V_m}{\text{norm}} \left(\phi_{vt} \sum_{k \in K} \tau_k \sum_{dx \in E_k} n_{kdx} \frac{\theta_{td}}{p(x|d)} \frac{\partial}{\partial \phi_{vt}} \prod_{u \in x} \phi_{ut} + \phi_{vt} \frac{\partial R}{\partial \phi_{vt}} \right) =$$

$$= \underset{v \in V_m}{\text{norm}} \left(\sum_{k \in K} \sum_{dx \in E_k} \tau_k n_{kdx} [v \in x] p_{tdx} + \phi_{vt} \frac{\partial R}{\partial \phi_{vt}} \right)$$

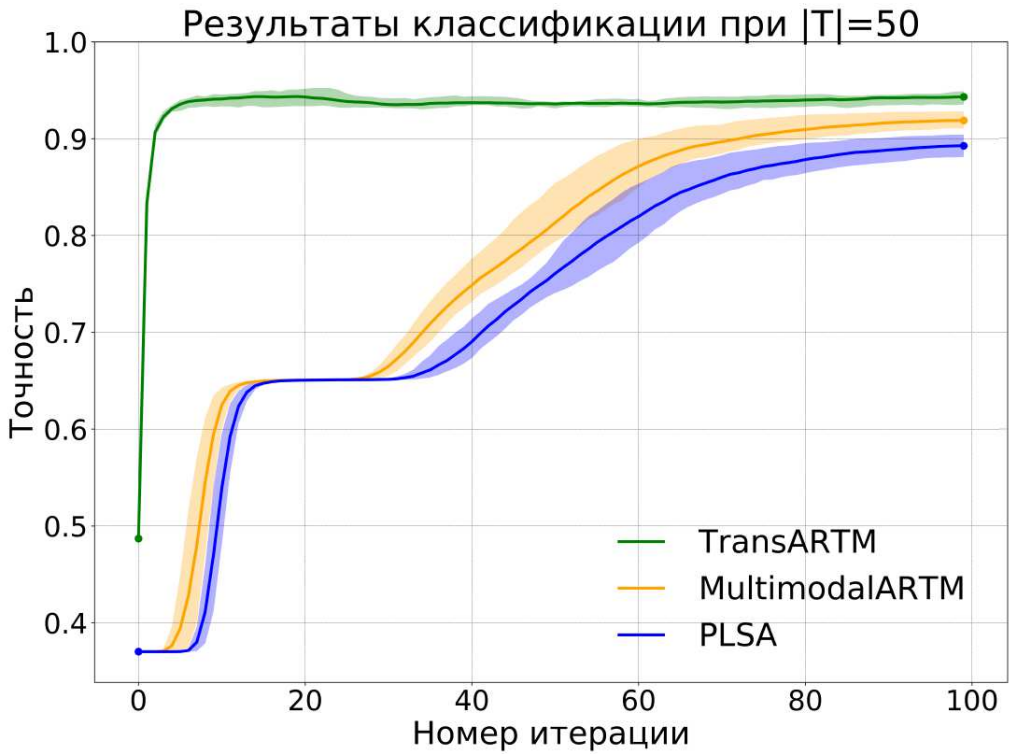
$$\theta_{td} = \underset{t \in T}{\text{norm}} \left(\theta_{td} \sum_{k \in K} \tau_k \sum_{x \in d} n_{kdx} \frac{1}{p(x|d)} \prod_{v \in x} \phi_{vt} + \theta_{td} \frac{\partial R}{\partial \theta_{td}} \right) =$$

$$= \underset{t \in T}{\text{norm}} \left(\sum_{k \in K} \sum_{x \in d} \tau_k n_{kdx} p_{tdx} + \theta_{td} \frac{\partial R}{\partial \theta_{td}} \right)$$

■

Эксперименты на модельных данных

13M транзакций, 3 модальности, 5 классов, 9 типов рёбер

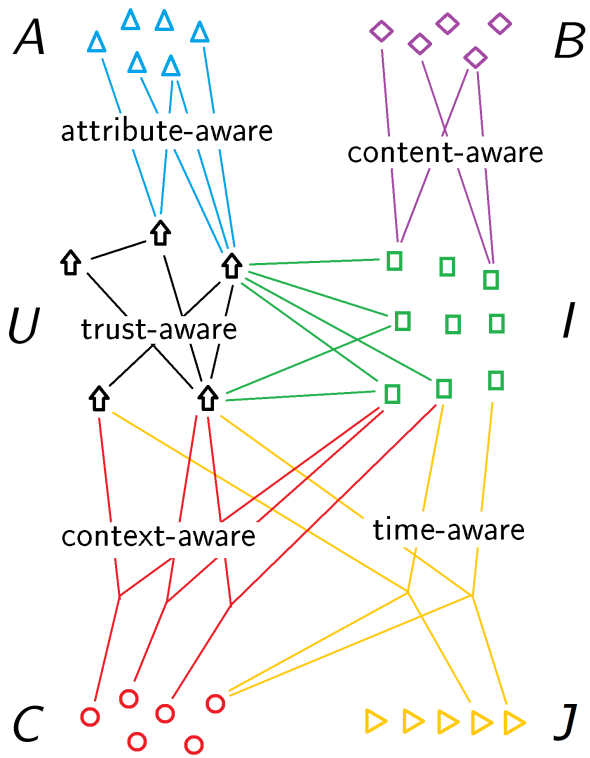


Вывод: обычные ТМ восстанавливают гиперграф плохо и долго

Илья Жариков. Гиперграфовые тематические модели транзакционных данных. Магистерская диссертация, МФТИ, 2018.

Транзакционные данные в рекомендательных системах

U — конечное множество (словарь) клиентов (users)
I — конечное множество (словарь) объектов (items)
A — словарь атрибутов клиентов (соцдем, регион, хобби...)
B — словарь свойств объектов (слова в текстовых объектах)
C — словарь ситуативных контекстов
J — словарь интервалов времени



Возможные виды данных:

n_{ui} — клиент u выбрал объект i
n_{ua} — клиент u имеет атрибут a
n_{ib} — объект i имеет свойство b
n_{uv} — клиент u доверяет клиенту v
n_{uib} — клиент u отметил i тэгом b
n_{uic} — клиент u выбрал i в контексте c
n_{uicj} — u выбрал i в c в интервале j

Симметризованная модель транзакционных данных

Симметризованные модели подходят для задач, в которых нет «естественных контейнеров» с неизменным содержимым

$x \subset V$ — рёбра гиперграфа
$\phi_{vt} = p(v|t)$ — распределение термов $v \in V^m$ в теме t
$\pi_t = p(t)$ — распределение тем во всей коллекции
E_k — наблюдаемая выборка рёбер-транзакций $x \subset V$ типа k
n_{kx} — число наблюдений ребра x в выборке E_k

Вероятностная тематическая модель рёбер гиперграфа:

$$p(x) = \sum_{t \in T} p(t) \prod_{v \in x} p(v|t) = \sum_{t \in T} \pi_t \prod_{v \in x} \phi_{vt}$$

Задача максимизации регуляризованного правдоподобия:

$$\sum_{k \in K} \tau_k \sum_{x \in E_k} n_{kx} \ln \left(\sum_{t \in T} \pi_t \prod_{v \in x} \phi_{vt} \right) + R(\Phi, \pi) \to \max_{\Phi, \pi}$$

EM-алгоритм для симметризованной гиперграфовой ARTM

Задача максимизации регуляризованного правдоподобия:

$$\sum_{k\in K}\tau_k\sum_{x\in E_k}n_{kx}\ln\left(\sum_{t\in T}\pi_t\prod_{v\in x}\phi_{vt}\right) + R(\Phi,\pi)\to\max_{\Phi,\pi}$$

EM-алгоритм: метод простой итерации для системы уравнений со вспомогательными переменными $p_{tx}=p(t|x)$:

$$\text{E-шаг:}\quad p_{tx}=\operatorname*{norm}_{t\in T}\left(\pi_t\prod_{v\in x}\phi_{vt}\right)$$

$$\text{M-шаг:}\quad \begin{cases}\phi_{vt}=\operatorname*{norm}_{v\in V^m}\left(\sum_{k\in K}\tau_k\sum_{x\in E_k}\big[v\in x\big]n_{kx}p_{tx}+\phi_{vt}\frac{\partial R}{\partial\phi_{vt}}\right)\\ \pi_t=\operatorname*{norm}_{t\in T}\left(\sum_{k\in K}\tau_k\sum_{x\in E_k}n_{kx}p_{tx}+\pi_t\frac{\partial R}{\partial\pi_t}\right)\end{cases}$$

Симметризованная модель рекомендательной системы

Сумма log-правдоподобий для четырёх типов транзакций (content-attribute-context-time-aware model):

$$\sum_{u,i} n_{ui} \ln \sum_{t \in T} \pi_t \, \phi_{ut} \, \phi_{it}$$
$$+ \tau_1 \sum_{i,b} n_{ib} \ln \sum_{t \in T} \pi_t \, \phi_{it} \, \phi_{bt}$$
$$+ \tau_2 \sum_{u,a} n_{ua} \ln \sum_{t \in T} \pi_t \, \phi_{ut} \, \phi_{at}$$
$$+ \tau_3 \sum_{u,i,c,j} n_{uicj} \ln \sum_{t \in T} \pi_t \, \phi_{it} \, \phi_{ut} \, \phi_{ct} \, \phi_{jt} \;\to\; \max_{\Phi,\pi}$$

Как построить симметризованную модель в BigARTM:

1. документы становятся модальностью
2. коллекция разбивается на документы d по времени
3. столбцы θ_{td} сглаживаются по n_t или по $\theta_{t,d-1}$

Модели предложений и коротких текстов TwitterLDA, senLDA

S_d — множество предложений документа d

n_{sw} — сколько раз терм w встречается в предложении s

Тематическая модель предложения s:

$$p(s|d) = \sum_{t \in T} p(t|d) \prod_{w \in s} p(w|t)^{n_{sw}} = \sum_{t \in T} \theta_{td} \prod_{w \in s} \phi_{wt}^{n_{sw}}$$

Максимизация регуляризованного log-правдоподобия

$$\sum_{d \in D} \sum_{s \in S_d} \ln \sum_{t \in T} \theta_{td} \prod_{w \in s} \phi_{wt}^{n_{sw}} + R(\Phi, \Theta) \to \max_{\Phi, \Theta}$$

это частный случай гиперграфовой модели, предложения являются гипер-рёбрами.

Wayne Xin Zhao, Jing Jiang, Jianshu Weng, Jing He, Ee Peng Lim et al. Comparing Twitter and traditional media using topic models. ECIR 2011.

G.Balikas, M.-R.Amini, M.Clausel. On a topic model for sentences. SIGIR 2016.

Гиперграфовые тематические модели языка

Гипер-рёбрами могут быть *сегментоиды* — подмножества термов, связанные по смыслу и порождаемые общей темой:

- предложение / фраза / синтагма
- ветка синтаксического дерева / именная группа
- факт «объект, субъект, действие»
- пары синонимов, гипоним–гипероним, мероним–холоним
- лексическая цепочка
- текст комментария, дата–время, автор

Модель даёт интерпретируемые тематические эмбединги:

- $p(t|d)$ — каждого контейнера, в частности, документа
- $p(t|w) = \phi_{wt}\frac{p(t)}{p(w)}$ — каждого терма, в частности, слова
- $p(t|d,x)$ — каждой отдельной транзакции (фразы, факта)

Анализ транзакций розничных клиентов банка

Дано (Sberbank Data Science Contest):

D — множество клиентов (15 000)

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

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

Найти: темы — типы экономического поведения (потребления)

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

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

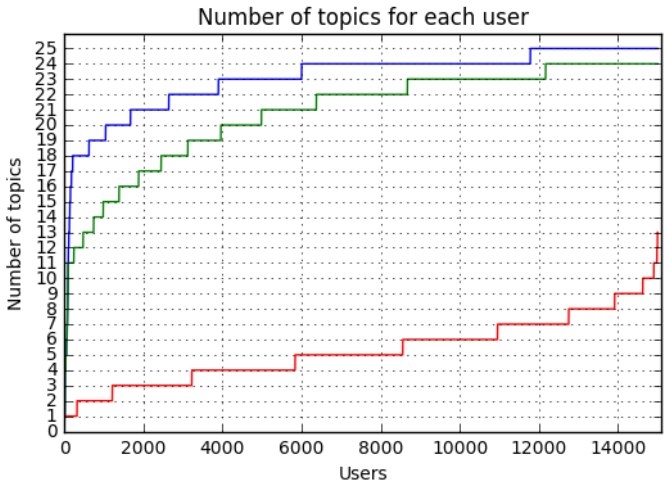
Регуляризаторы:

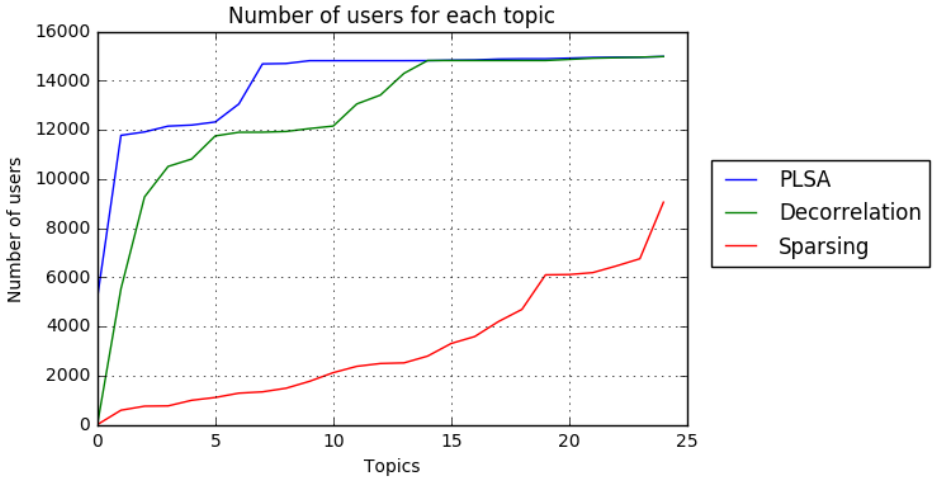
- повышение различности тем
- разреживание $p(t|d)$
- учёт модальностей времени, типа транзакции, терминала

Egorov E., Nikitin F., Goncharov A., Alekseev V., Vorontsov K. Topic Modelling for Extracting Behavioral Patterns from Transactions Data // IC-AIAI 2019.

Построение модели ARTM, 25 тем

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





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

Пользуюсь картой только чтобы снять наличные

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

Наличные + авто, спорт, компьютеры

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

Цивилизованный потребитель: разные магазины, связь, авто

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

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

Продвинутые мамки

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

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

Бизнес-леди: забыла про наличку — всё по карте

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

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

Продвинутый активный потребитель всего, и по карте

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

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

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

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

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

Провинциальный малый бизнес

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

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

Анализ транзакций корпоративных клиентов банка

Данные:
лесная отрасль, 2016 г., 10.7M транзакций, 1M компаний.
Транзакция — это тройка ⟨покупатель, продавец, текст⟩.

Некоторые *тексты* платёжных поручений (далеко не все!) содержат названия товаров и услуг.

Документ — это история транзакций одной компании

Семь модальностей:

- компании: поставщики / покупатели
- слова в платёжных поручениях: поставщики / покупатели
- ОКВЭДы данной компании
- ОКВЭДы контрагентов: поставщики / покупатели

Примеры тем — видов деятельности компаний

покупка	продажа
0.11: услуга	0.12: лдсп
0.07: классик	0.08: дсп
0.05: дрова	0.03: мдф
0.05: пиловочник	0.03: поставка
0.05: материал	0.02: услуга
0.03: порода	0.02: охранный
0.03: лесоматериал	0.02: ламинировать
0.03: сертум	0.02: хдф
0.02: хвойный	0.02: материал
0.01: дерево	0.01: накл
0.01: транспортный	0.01: товар

покупка	продажа
0.19: право	0.16: арендный
0.17: сбис	0.10: часть
0.16: использование	0.08: плата
0.03: аккаунт	0.04: минимальный
0.02: электронный	0.04: участок
0.02: лицевой	0.04: использование
0.02: устный	0.02: земля
0.01: устройство	0.02: лесничество
0.01: генерация	0.02: земельный
0.01: хранение	0.01: фонд
0.01: ключевой	0.01: федеральный

Примеры тем — видов деятельности компаний

покупка		продажа
0.09: ткань		0.16: мебель
0.09: поставка		0.05: плёнка
0.02: мебельный		0.04: стул
0.02: деревянный		0.03: кресло
0.02: транспортный		0.03: изделие
0.02: фанера		0.02: краска
0.02: поролон		0.02: фанера
0.01: механизм		0.01: лкм
0.01: плата		0.01: лакокрасочный
0.01: частичный		0.01: лак
		0.01: материал
		0.01: клеить

покупка		продажа
0.06: лдсп		0.37: товар
0.05: фурнитура		0.15: мебель
0.02: плёнка		0.04: поставка
0.02: материал		0.04: накладный
0.02: мебельный		0.03: накл
0.02: стекло		0.03: рубль
0.02: мдф		
0.02: кромка		
0.01: транспортный		
0.01: клеить		
0.01: профиль		
0.01: пвх		

Примеры тем — видов деятельности компаний

покупка		продажа
0.52: гсм		0.14: вывоз
0.43: далее		0.09: тбо
		0.04: мусор
		0.03: отход
		0.02: утилизация
		0.01: тко

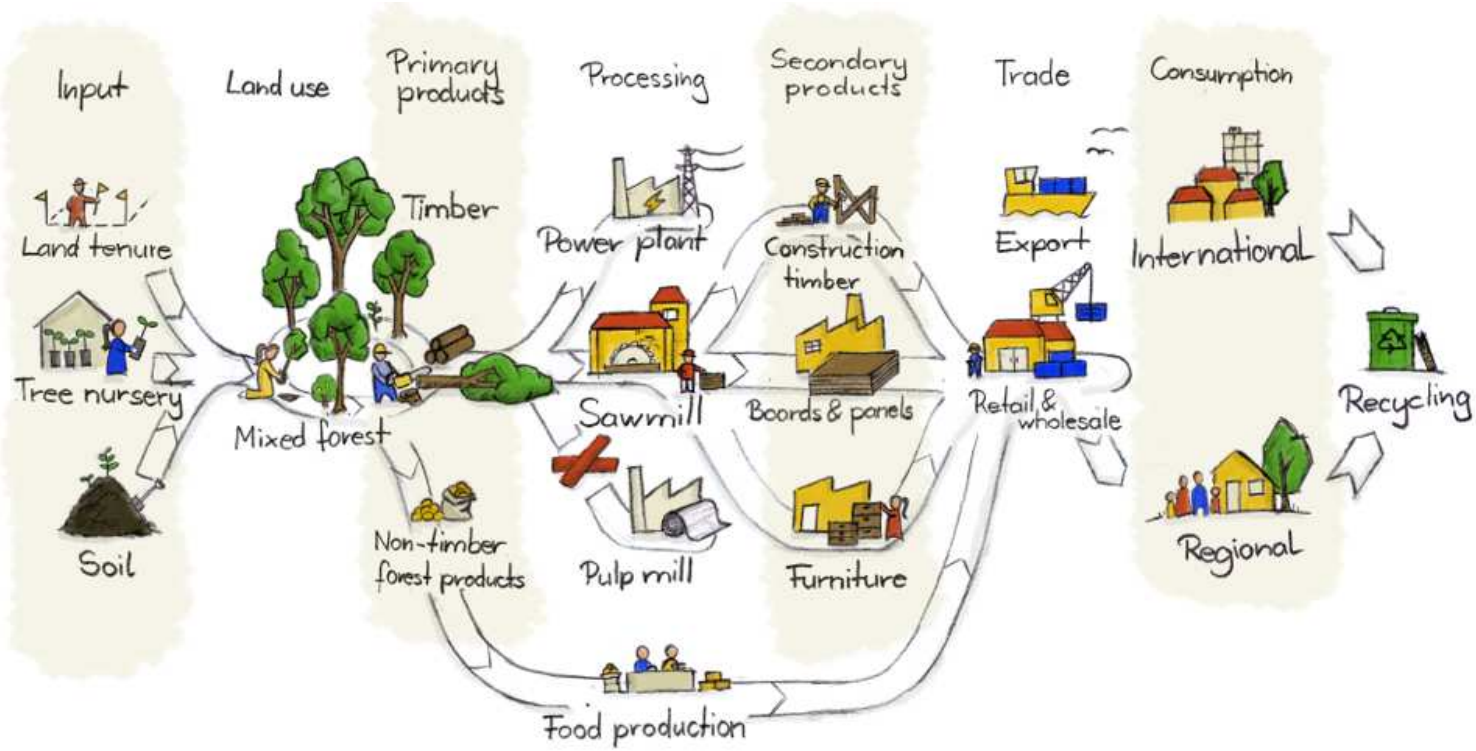
покупка		продажа
0.19: налог		0.11: бумага
0.06: услуга		0.08: гофроящик
0.04: макулатура		0.04: гофрокартон
0.03: поставка		0.03: гофрокороб
0.03: транспортный		0.03: поставка
0.02: лесопродукция		0.03: фактура
0.02: автоуслуга		0.02: гофропродукция
0.01: перевозка		0.02: гофротару
0.01: плата		0.02: гофрирование
		0.02: гофролоток
		0.02: товар
		0.01: лоток

Примеры тем — видов деятельности компаний

покупка	продажа	продажа	продажа
0.15: программа	0.13: фурнитура	0.14: рекламный	0.21: тмц
0.11: право	0.09: материал	0.13: размещение	0.06: накл
0.09: сертификат	0.08: лдсп	0.09: материал	0.04: инструмент
0.07: эвм	0.04: кромка	0.05: проект	0.03: пила
0.07: использование	0.04: мебельный	0.05: яндекс	0.02: заточка
0.07: лицензия	0.04: фрз	0.04: директ	0.02: нож
0.04: криптопро	0.04: мдф	0.04: реклама	0.02: материал
0.03: абонентский	0.03: клеить	0.02: рубль	0.02: фреза
0.02: обслужа	0.03: пвх	0.01: стек	0.02: клеить
0.02: пользование	0.02: тмц		0.01: товар
0.02: контур	0.02: комплект		0.01: перчатка
0.01: проверка	0.02: профиль		
	0.02: столешница		

Конечные цели моделирования транзакционных данных

- Получение векторных представлений компаний
- Поиск схожих и конкурирующих компаний
- Восстановление структуры товарных потоков отрасли



Резюме: тематические модели связного текста

Пять механизмов моделирования (от сильного к слабому):

1. *Линейная тематизация с локальным E-шагом* (лекция 4): вместо тематики документа $p(t|d)$ вычисляется тематика локального контекста $p(t|i)$ и $p(t|d, w_i)$, $i = 1, \ldots, n_d$.
2. *Пост-обработка E-шага* (сегодня): тематические векторы $p(t|d, w_i)$ подвергаются эвристическим преобразованиям, что эквивалентно регуляризации E-шага.
3. *Гиперграфовая тематическая модель* (сегодня): термы, порождаемые общей темой (предложения, фразы, факты, и т. д.), объединяются в «транзакции».
4. *Тематическая модель сети слов или битермов* (лекция 2): по каждому слову формируется псевдодокумент путём объединения (в «мешок») всех его локальных контекстов.
5. *Тематическая модель n-грамм* (лекции 2 & 6): n-граммы выделяются заранее и используются как модальность, для сокращения словарей оценивается тематичность термов.

Задания по курсу

Задача-минимум: научиться решать задачи анализа текстов с использованием тематического моделирования

Задача-максимум: получить новый научный результат

виды деятельности	оценка
теоретическая задача	X
теоретическая задача*	$2X$
теоретическая задача**	$3X$
решение прикладной задачи	$10X$
обзор по последним PTM/NTM	$10X$
участие в проекте	$20X$
работа над открытой проблемой	$25X$

где X — оценка за вид деятельности по 5-балльной шкале.
score — суммарная оценка по всем видам деятельности.

Итоговая оценка: $\min\left(5, \lfloor \text{score}/20 \rfloor\right)$ по 5-балльной шкале.

Задания к лекции 1

Упражнения на принцип максимума правдоподобия:

1. Биграммная модель коллекции: $p(w|v) = \xi_{wv}$,
где v — слово, идущее в тексте перед w.
Найти параметры модели ξ_{wv}.

2. Биграммная модель документов: $p(w|v,d) = \xi_{dwv}$.
Найти параметры модели ξ_{dwv}.

Подсказка: применить условия ККТ или основную лемму.

3*. Творческое задание (возможны разные решения).
Предложите модель, разделяющую роли слов в текстах:
— тематические слова
— специфичные слова документа (шум)
— слова общей лексики (фон)

Подсказка 1: искать распределение ролей слов $p(r|w)$, $r \in \{\text{т}, \text{ш}, \text{ф}\}$.
Подсказка 2: можно разреживать $p(r|w)$ для жёсткого определения ролей.
Подсказка 3: можно использовать документную частоту слов.

Задания к лекции 2

4. Пользуясь основной леммой, докажите, что регуляризатор битермов эквивалентен добавлению псевдодокументов d_u в исходную коллекцию (см. слайд 13)

Прикладная исследовательская задача:
автоматическое выделение научных терминов (ATE)

- Дано:
  коллекция размеченных текстов конкурса ruTermEval;
  неразмеченная коллекция текстов той же тематики
- Найти:
  метод ATE на основе комбинирования ARTM и TopMine;
  обоснование, что синтаксический анализ не нужен;
  зависимость качества ATE от объёма коллекции
- Критерий:
  качество ATE (Prec, Rec, F1) на размеченных данных

Теоретическое задание к лекции 3

Выведете EM-алгоритм для тематической языковой модели:

5. $p(w|d) = \sum_t \phi_{wt}\theta_{td}$, используя в качестве исходных данных последовательность $(d_i, w_i)_{i=1}^n$ вместо счётчиков n_{dw}.
Докажите эквивалентность обычному EM-алгоритму ARTM.

6. $p(w|d) = \sum_t \phi_{tw}\frac{p(w)}{p(t)}\theta_{td}$, где $p(t)$ фиксировано,
$\phi_{tw} = p(t|w)$, $\theta_{td} = p(t|d)$ — параметры модели.

7. $p(w|d) = \sum_t \phi_{tw}\frac{p(w)}{p(t)}\theta_{td}$, где $p(t)$ фиксировано,
$\phi_{tw} = p(t|w)$ — параметры модели, $\theta_{td} = \sum_w \frac{n_{dw}}{n_d}\phi_{tw}$.

8*. Фиксация $p(t)$ как внешнего параметра упрощает выкладки, но может нарушать условия целостности модели:

$$p(t) = \sum_t \phi_{tw}p(w), \qquad p(t) = \sum_t \theta_{td}p(d).$$

Как обеспечить выполнение этих условий в EM-алгоритме?

Теоретическое задание к лекции 4

9. Докажите, что необходимым условием максимума

$$\sum_{i=1}^{n} \ln \sum_{t \in T} p(w_i, t|i, \Omega) \;\to\; \max_{\Omega}$$

для языковой модели со скрытыми переменными $t \in T$ (не обязательно темами) и параметрами $\Omega = (\omega_{kj})$ — набором неотрицательных нормированных векторов, является система

$$\text{E-шаг:} \quad p(t|w_i, i) = \operatorname*{norm}_{t \in T} p(w_i, t|i, \Omega)$$

$$\text{M-шаг:} \quad \omega_{kj} = \operatorname*{norm}_{k} \left(\sum_{i=1}^{n} \sum_{t \in T} p(t|w_i, i)\, \omega_{kj} \frac{\partial}{\partial \omega_{kj}} \ln p(w_i, t|i, \Omega) \right)$$

10. Выведете отсюда EM-алгоритм для частных случаев:

1) $p(w, t|i, \Omega) = \phi_{wt}\theta_{td_i}$;
2) $p(w, t|i, \Omega) = \phi_{tw} \frac{p(w)}{p(t)} \sum_{w \in d_i} \frac{n_{d_i w}}{n_{d_i}} \phi_{tw}$;
3) $p(w, t|i, \Omega) = \phi_{tw} \frac{p(w)}{p(t)} \sum_{c \in C_i} \alpha_{ci} \phi_{tw_c}$.

11. Творческое задание.** Предложите способ ввести обучаемые параметры в тематическую модель внимания.

Исследовательская задача к лекции 4

Реализуйте EM-алгоритм для модели локального контекста (или воспользуйтесь готовой реализацией)

Исследуйте зависимость метрик качества модели

- перплексия: $\mathscr{P} = \exp\left(-\frac{1}{n}\sum_{i=1}^{n} p(w|C_i)\right)$
- разреженность, различность, когерентность тем
- дефекты целостности модели:
  $\left\|p(t) - \frac{n_t}{n}\right\|$, $\left\|p(t) - \sum_t \phi_{tw} p(w)\right\|$, $\left\|p(t) - \sum_t \theta_{td} p(d)\right\|$

от номера итерации и от параметров модели:

- $|T|$ — число тем
- L — число проходов
- τ — вес N_{tw} в формуле M-шага, особый случай $\tau = 0$
- $\vec{\gamma}_i$, $\overleftarrow{\gamma}_i$ — длина скользящего среднего
- $\vec{\gamma}_i$, $\overleftarrow{\gamma}_i$, β — баланс левого и правого контекста
- $\vec{\gamma}_i$, $\overleftarrow{\gamma}_i$ — учёт границ предложений, абзацев, секций
- опция «$i \in C_i$ или $i \notin C_i$»

Теоретическое задание к лекции 5

12. Найдите дискретное распределение $P = (p_i)_{i=1}^n$ в задаче $\sum_i n_i \mu(p_i) \to \max$ с гладкой монотонно возрастающей $\mu(p)$.
Отдельно рассмотрите случаи $\mu(p) = p^s$, $s = 1$, $s \to 0$.

13. Выведете EM-алгоритм в случае, когда ln заменён гладкой монотонно возрастающей функцией μ:

$$\sum_{d \in D} \sum_{w \in d} n_{dw} \mu\left(\sum_{t \in T} \phi_{wt} \theta_{td} \right) + R(\Phi, \Theta) \to \max_{\Phi, \Theta}$$

Подумайте, какие замены логарифма полезны, и почему.

14. Простейшая идея разреживания — обнуление малых вероятностей. Чтобы обосновать эту эвристику, найдите, какому регуляризатору соответствует формула M-шага

$$\phi_{wt} = \operatorname*{norm}_w \left(n_{wt} [n_{wt} > \gamma n_t] \right)$$

Подсказка: с учётом подстановки несмещённой оценки ϕ^*_{wt}

Теоретическое задание к лекции 5

Проект «Тематизатор». Аналитик построил модель $\Phi^0\Theta^0$ и отметил среди столбцов матрицы Φ^0 темы двух типов: удачные $T_+ \subset T$ и неудачные $T_- \subset T$.

Теперь он хочет построить модель ещё раз так, чтобы

- удачные темы остались в матрице Φ;
- остальные темы построились по-другому и были не похожи на каждую из неудачных тем $t \in T_-$.

15. Предложите регуляризаторы для этого.

16. Не получится ли так, что новые темы будут отдаляться от суммы неудачных тем $\sum_{t \in T_-} \phi^0_{wt}$ вместо того, чтобы отдаляться от каждой из неудачных тем по отдельности? Почему это плохо и как этого избежать?

17. Предложите способ инициализации Φ для новой модели.

Исследовательская задача к лекции 5

Продолжение исследования по автоматическому выделению научных терминов (Automatic Term Extraction, ATE)

- Дано:
  коллекция размеченных текстов конкурса ruTermEval;
  неразмеченная коллекция текстов той же тематики
- Найти:
  — оптимальную стратегию регуляризации на основе декоррелирования и сглаживания фоновых тем
  — рекомендации по управлению относительными коэффициентами регуляризации
  — критерий тематичности терминов по расстоянию между распределениями $p(t|w)$ и $p_0(t) = \frac{1}{|T|}$, позволяющий наиболее чётко отличать термины от фоновой лексики
- Критерий:
  — максимум доли терминов в предметных темах
  — минимум доли терминов в фоновых темах

Исследовательская задача к лекции 5

Продолжение исследования модели локального контекста (можно воспользоваться готовой реализацией EM-алгоритма)

Исследуйте устойчивость модели в сравнении с ARTM

- без регуляризации
- с регуляризатором декоррелирования, при различных значениях относительного коэффициента регуляризации

Как на устойчивость модели влияют её параметры:

- $|T|$ — число тем
- L — число проходов
- τ — вес N_{tw} в формуле M-шага, особый случай $\tau = 0$
- $\vec{\gamma}_i$, $\overleftarrow{\gamma}_i$ — длина скользящего среднего
- $\vec{\gamma}_i$, $\overleftarrow{\gamma}_i$, β — баланс левого и правого контекста
- $\vec{\gamma}_i$, $\overleftarrow{\gamma}_i$ — учёт границ предложений, абзацев, секций
- опция «$i \in C_i$ или $i \notin C_i$»

Теоретическое задание к лекции 6

18. Для иерархической тематической модели с рег. $R(\Phi, \Psi)$ предложите способ разреживания матрицы связей $\Psi = \big(p(s|t)\big)$, гарантирующий, что
1) у каждой родительской темы будет хотя бы одна дочерняя;
2) у каждой дочерней темы будет хотя бы одна родительская.

Подсказка: можно придумывать критерий регуляризации, а можно — формулу М-шага для матрицы Ψ.

19. Предложите способ гарантировать, что если родительская тема t получает только одну дочернюю s, то она переходит в неё целиком и как распределение: $p(w|s) = p(w|t)$, то есть тема t на данном уровне не расщепляется на подтемы.

20. Предложите способ согласования вероятностных смесей $p(w|t) \approx \sum_{s \in S} p(w|s)p(s|t)$ и $p(t|d) \approx \sum_{s \in S} p(t|s)p(s|d)$ с учётом тождества $p(s|t)p(t) = p(t|s)p(s)$.

Исследовательское задание к лекции 6

Проект «Мастерская знаний». Нужна тематическая модель подборок научных статей и/или поисковой выдачи.

Дано:

- 1000 подборок, в каждой по 1000 аннотаций научных статей, ранжированные по сходству с аннотацией-запросом по эмбедингам модели SciRus (эмбединги тоже даны)

Найти:

- метод согласования тематической модели с эмбедингами
- метод выделения терминов (Automatic Term Extraction)
- метод отбора терминов по тематичности
- метод отсева тематически нерелевантных аннотаций

Критерии:

- согласованность тематической модели с эмбедингами
- интерпретируемость тем
- качество выделения терминов

Теоретическое задание к лекции 7

21. Выведите EM-алгоритм с регуляризатором семантической однородности, предполагая, что n_{tdw} и n_t — константы (внешние параметры, не зависящие от Φ, Θ).
Докажите, что подстановка этого регуляризатора в M-шаг эквивалентна введению мультипликативной поправки $(1 + \tau\beta_{dw})$ в критерий log-правдоподобия.

22.** Выведите EM-алгоритм с регуляризатором семантической однородности, предполагая, что n_{tdw} и n_t выражаются через параметры модели Φ, Θ.

23*. Предложите формулу средневзешенных статистик S_* для тематической модели локальных контекстов.
Проверьте, что полученная формула совпадает с введённой на лекции, если контекстом является весь документ.

Исследовательская задача к лекции 7

Исследование EM-алгоритма для модели локального контекста

- Оценивание внутритекстовой когерентности
  - реализуйте вычисление средневзвешенной когерентности
  - подберите наилучшее сочетание эвристик rel и coh в калибровочном эксперименте без экспертной разметки
  - какие эвристики в модели локального контекста улучшают внутритекстовую когерентность?
  - воспроизводимо ли это улучшение на разных коллекциях?
- Оценивание средневзвешенных статистик
  - реализуйте вычисление S_t, S_{wt}
  - как зависит вид распределения $\{S_t\}$ от числа тем?
  - есть ли корреляция между S_t и когерентностью coh_t?
  - предложите способ разделения темы с большим S_t на подтемы и их инициализацию термами с большими S_{wt}
- Оценивание несбалансированности тем
  - реализуйте генератор коллекций с заданным дисбалансом тем
  - как дисбаланс влияет на число разделённых и слитых тем?
  - модели локального контекста лишены этой проблемы?
  - уменьшает ли регуляризатор семантической однородности число разделённых и слитых тем?

Примеры датасетов для практических заданий по курсу

1. Открытые датасеты (английский): 20NG, NIPS, KOS
2. Ранжированные результаты поиска научных статей (по данным eLibrary, arXiv, PubMed)
3. Научно-популярные статьи: ПостНаука, Элементы, Хабр,...
4. Техноблоги: Хабр (русский), TechCrunch (английский)
5. Данные социальных сетей: VK, Twitter, Telegram,...
6. Статьи по Complexity Sciences (для хронокарты науки)

- Википедия
- Викиновости (1.5M статей, проект закрыт 30/03/2026)
- Данные кадровых агентств: резюме + вакансии
- Транзакции клиентов Sberbank DSD 2016
- Акты арбитражных судов РФ

---

http://bigartm.org
http://drive.google.com/drive/folders/1PPnw6aZOJAJoLRYuwdGm437RssV-XQx0

Проекты

- «Тематизатор» для социо-гуманитарных исследований:
  — пользователь задаёт грубый фильтр текстового потока;
  — задача: «классифицировать иголки в стоге сена»,
  — разделив темы на информативные и мусорные,
  — выделив аспекты и тональности в каждой теме;
  — конечная цель: кол./кач. анализ предметной области,
  — реализация данного сценария как модуля в среде Orange

- «Мастерская знаний» для научного поиска:
  — пользователь строит тематические подборки статей,
  — поисковая выдача формируется моделью SciRus;
  — задача: показать пользователю тематику подборки;
  — понадобится: автоматическое выделение терминов,
  — выделение тематических фраз из документов,
  — автоматическое именование и суммаризация тем;
  — конечная цель: помочь в понимании предметной области

Открытые проблемы тематического моделирования

1. Тематические модели внимания последовательного текста
2. Проблема несбалансированности тем в коллекции
3. Измерение интерпретируемости тем (когерентность)
4. Обеспечение 100%-й интерпретируемости тем
5. Автоматическое именование и суммаризация тем
6. Калибровка моделей тематической фильтрации
7. Согласование тем с предобученными эбмедингами LLM
8. Статистические оценки состоятельности тем
9. Обнаружение новых тем или трендов в потоке текстов
10. Обеспечение устойчивости и полноты множества тем
11. Автоматический подбор гиперпараметров, AutoML
12. Гиперграфовые тематические модели для RecSys