

# Вероятностные тематические модели

## Лекция 6. Регуляризаторы для АРТМ

К. В. Воронцов  
vokov@forecsys.ru

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

ВМК МГУ • весна 2016

## 1 Корреляции и связи

- Регрессия и классификация на текстах
- Модель СТМ (Correlated Topic Model)
- Гиперссылки, цитирование, влияние

## 2 Время и пространство

- Регуляризаторы времени
- Эксперименты на коллекции пресс-релизов
- Гео-пространственные модели

## 3 Социальные сети

- Тематические сообщества (topical community)
- Направленные связи
- Социальные роли

## Напоминание. Задача тематического моделирования

**Дано:**  $W^m$  — словарь токенов  $m$ -й модальности,  $m \in M$ ,  
 $D$  — коллекция текстовых документов  $d \subset W = \bigsqcup_m W^m$ ,  
 $n_{dw}$  — сколько раз токен  $w$  встретился в документе  $d$ .

**Найти:** модель  $p(w|d) = \sum_{t \in T} \phi_{wt} \theta_{td}$  с параметрами  $\Phi_{W^m \times T}^m$  и  $\Theta_{T \times D}$

$\phi_{wt} = p(w|t)$  — вероятности токена  $w$  в каждой теме  $t$ ,

$\theta_{td} = p(t|d)$  — вероятности тем  $t$  в каждом документе  $d$ .

**Критерий** максимума регуляризованного правдоподобия:

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

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

## Напоминание. Регуляризованный EM-алгоритм

Максимизация  $\ln$  правдоподобия с регуляризатором  $R$ :

$$\sum_{d \in D} \sum_{w \in d} \tilde{n}_{dw} \ln \sum_{t \in T} \phi_{wt} \theta_{td} + R(\Phi, \Theta) \rightarrow \max_{\phi, \theta}$$

где  $\tilde{n}_{dw} = \tau_{m(w)} n_{dw}$ ,  $m(w)$  — модальность токена  $w$ .

EM-алгоритм: метод простой итерации для системы уравнений

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

## Регуляризатор для классификации и категоризации текстов

**Цель:** построить тематическую модель классификации.

$C$  — множество классов (категорий);

$n_{dc} = [\text{документ } d \text{ относится к классу } c]$  — обучающие данные;

$p(c|d) = \sum_{t \in T} \phi_{ct} \theta_{td}$  — линейная модель классификации.

Регуляризатор — правдоподобие модальности классов:

$$R(\Phi, \Theta) = \tau \sum_{d \in D} \sum_{c \in C} n_{dc} \ln \sum_{t \in T} \phi_{ct} \theta_{td} \rightarrow \max,$$

это тематическая модель с двумя модальностями,  $W$  и  $C$ .

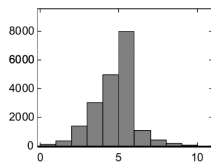
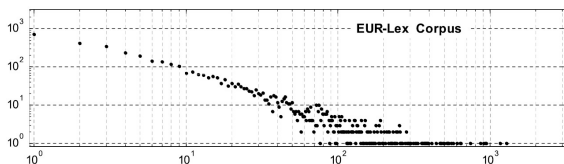
ТМ превосходит SVM в случае несбалансированных классов.

---

*Rubin T. N., Chambers A., Smyth P., Steyvers M.* Statistical topic models for multi-label document classification // Machine Learning, 2012, no. 1–2.

## Эксперимент. Категоризация коллекции EUR-Lex

- EUR-Lex:  $|D| = 19\,800$  документов — законы Евросоюза
- Две модальности:  $W^1$  слова (21К),  $W^2$  категории (3 250)
- Категории несбалансированные и пересекающиеся:



- слева:  $\#$  категорий с заданным  $\#$  документов в категории
- справа:  $\#$  документов с заданным  $\#$  категорий

Rubin T. N., Chambers A., Smyth P., Steyvers M. Statistical topic models for multi-label document classification. Machine Learning, 2012, 88(1-2), p. 157–208.

## Эксперимент. Категоризация коллекции EUR-Lex

### Регуляризаторы:

- Равномерное сглаживание  $\Theta$
- Равномерное сглаживание матрицы слова–темы  $\Phi^1$
- *Label regularization* для матрицы категории–темы  $\Phi^2$ :

$$R(\Phi^2) = \tau \sum_{c \in W^2} \hat{p}_c \ln p(c) \rightarrow \max,$$

где  $p(c) = \sum_{t \in T} \phi_{ct} p(t)$  — распределение на категориях  $c$ ,

$p(t) = \frac{n_t}{n}$  — распределение на темах,

$\hat{p}_c$  — доля документов категории  $c$  в обучающей выборке.

---

*Mann G. S., McCallum A.* Simple, robust, scalable semi-supervised learning via expectation regularization // ICML 2007, Pp. 593–600.

## Эксперимент. Категоризация коллекции EUR-Lex

DLDA (Dependency LDA) [Rubin 2012] — ближайший аналог ARTM для задач классификации среди байесовских моделей

Критерии качества [Rubin 2012]:

- AUC-PR (% ,  $\uparrow$ ) — Area under precision-recall curve
- AUC (% ,  $\uparrow$ ) — Area under ROC curve
- OneErr (% ,  $\downarrow$ ) — One error (most ranked label is not relevant)
- IsErr (% ,  $\downarrow$ ) — Is error (no perfect classification)

Результаты сравнения:

	AUC-PR	AUC	OneErr	IsErr
BigARTM	<b>51.3</b>	98.0	<b>29.1</b>	<b>95.5</b>
DLDA [Rubin 2012]	49.2	<b>98.2</b>	32.0	97.2
SVM	43.5	97.5	31.6	98.1



## Регуляризатор для задач регрессии

**Цель:** построить тематическую модель регрессии.

$y_d \in \mathbb{R}$  для всех документов  $d \in D$  — обучающие данные.

$E(y|d) = \sum_{t \in T} v_t \theta_{td}$  — линейная модель регрессии,  $v \in \mathbb{R}^{|T|}$ .

Регуляризатор — среднеквадратичная ошибка (МНК):

$$R(\Theta, v) = -\tau \sum_{d \in D} \left( y_d - \sum_{t \in T} v_t \theta_{td} \right)^2 \rightarrow \max$$

Подставляем, получаем формулы М-шага:

$$\theta_{td} = \text{norm}_t \left( n_{td} + \tau \left( y_d - \sum_{t \in T} v_t \theta_{td} \right) \right);$$

$$v = (\Theta \Theta^T)^{-1} \Theta y.$$

---

*Sokolov E., Bogolubsky L. Topic Models Regularization and Initialization for Regression Problems // CIKM-2015 Workshop on Topic Models. ACM, pp. 21–27.*

## Примеры задач регрессии на текстах

### MovieReview [Pang, Lee, 2005]

$d$  — текст отзыва на фильм

$y_d$  — рейтинг фильма (1..5), поставленный автором отзыва

### Salary (kaggle.com: *Adzuna Job Salary Prediction*)

$d$  — описание вакансии, предлагаемой работодателем

$y_d$  — годовая зарплата

### Yelp (kaggle.com: *Yelp Recruiting Competition*)

$d$  — отзыв (на ресторан, отель, сервис и т.п.)

$y_d$  — число голосов «useful», которые получит отзыв

---

*B. Pang, L. Lee. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales // ACL, 2005.*



## Обобщённое лог-нормальное распределение

**Мотивация.** Темы могут коррелировать. «Статьи по геологии чаще связаны с археологией, чем с генетикой».

**Гипотеза.** Вектор-столбцы  $\theta_d$  порождаются  $|T|$ -мерным лог-нормальным распределением с ковариационной матрицей  $S$ :

$$(\theta_{td}) = \text{SoftMax}(\eta_{td}) = \frac{\exp(\eta_{td})}{\sum_{s \in T} \exp(\eta_{sd})};$$

$$p(\eta_d | \mu, S) = \frac{1}{(2\pi)^{\frac{n}{2}} |S|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\eta_d - \mu)^T S^{-1}(\eta_d - \mu)\right),$$

$\mu, S$  — параметры гауссовского распределения.

$\eta_{td} = \ln \theta_{td} + C_d$  — векторы документов, определённые с точностью до константы  $C_d$ , не зависящей от тем  $t$ .

## Регуляризатор коррелированной тематической модели СТМ

Максимизация правдоподобия выборки векторов  $\eta_d = (\eta_{td})$ :

$$\sum_{d \in D} \ln p(\eta_d | \mu, S) \rightarrow \max.$$

Регуляризатор с параметрами  $\mu, S$ :

$$R(\Theta) = -\frac{\tau}{2} \sum_{d \in D} (\eta_d - \mu)^\top S^{-1} (\eta_d - \mu) \rightarrow \max.$$

Формулы M-шага ( $S, \mu$  можно обновлять намного реже, чем  $\Theta$ ):

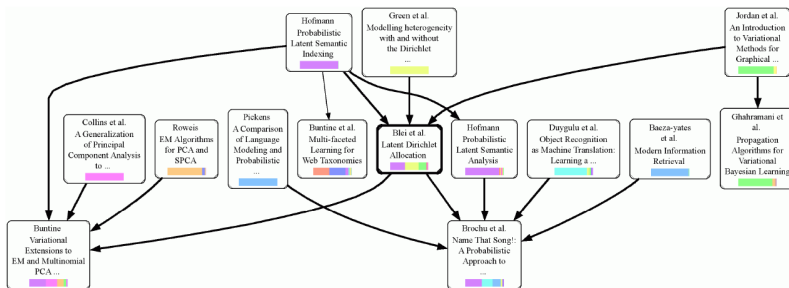
$$\theta_{td} = \operatorname{norm}_{t \in T} \left( n_{td} - \tau \sum_{s \in T} S_{ts}^{-1} (\ln \theta_{sd} - \mu_s) \right);$$

$$\mu = \frac{1}{|D|} \sum_{d \in D} \ln \theta_d;$$

$$S = \frac{1}{|D|} \sum_{d \in D} (\ln \theta_d - \mu) (\ln \theta_d - \mu)^\top.$$

## Модели, учитывающие цитирования или гиперссылки

- Учёт ссылок уточняет тематическую модель
- Тематическая модель выявляет влиятельные ссылки



Laura Dietz, Steffen Bickel, Tobias Scheffer. Unsupervised prediction of citation influences // ICML-2007, Pp. 233–240.

## Регуляризатор $\Theta$ для учёта связей между документами

**Цель:** улучшить темы, используя ссылки или цитирования (если документы ссылаются друг на друга, то их темы близки):

$n_{dc}$  — число ссылок из  $d$  на  $c$ .

Максимизируем ковариации связанных документов  $\theta_d, \theta_c$ :

$$R(\Theta) = \tau \sum_{d,c \in D} n_{dc} \text{cov}(\theta_d, \theta_c) \rightarrow \max.$$

Подставляем, получаем ещё один вариант сглаживания:

$$\theta_{td} = \text{norm}_t \left( n_{td} + \tau \theta_{td} \sum_{c \in D} n_{dc} \theta_{tc} \right).$$

---

*Laura Dietz, Steffen Bickel, Tobias Scheffer.* Unsupervised prediction of citation influences // ICML 2007. — Pp. 233–240.

## Связи как модальность. Регуляризатор $\Phi$

**Проблема учёта связей** в онлайн-овом EM-алгоритме:  
связанные документы могут оказаться в разных пакетах.

Документы содержат слова  $w \in W^1$  и ссылки  $c \in W^2 \subseteq D$   
 $W^2$  — модальность документов, на которые есть ссылки

Регуляризатор — log-правдоподобие модальности  $W^2$ :

$$R(\Phi^2, \Theta) = \tau \sum_{d \in D} \sum_{c \in W^2} n_{dc} \ln \sum_{t \in T} \phi_{ct} \theta_{td} \rightarrow \max.$$

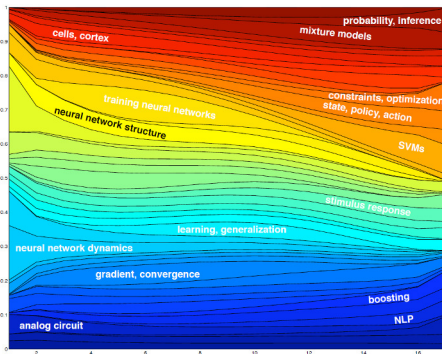
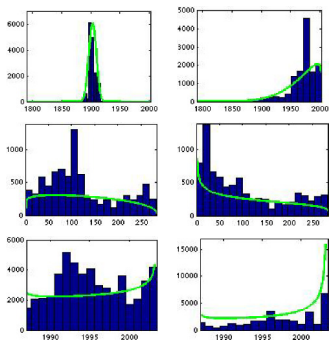
Другой вариант — сумма ковариационных регуляризаторов:

$$R(\Phi^2, \Theta) = \tau \sum_{d,c} n_{dc} \sum_{t \in T} \phi_{ct} \theta_{td} \rightarrow \max.$$



## Модель TOT (Topics over Time)

1. Каждая тема имеет непрерывное  $\beta$ -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

### Наши предположения:

- Время  $y \in Y$  дискретно, поскольку
  - меток времени в документах конечное число,
  - накопление статистики по темам требует времени.
- Сами темы  $p(w|t)$  не меняются во времени, меняются только их доли  $p(t|y)$ .
- Невозможно адекватно описать временной ряд  $p(y|t)$  параметрической моделью типа  $\beta$ -распределения.

### Цели моделирования:

- Выделить событийные и перманентные темы.
- Проследить развитие тем во времени.
- Выделить тренды (в новостях, в научных публикациях).

Регуляризаторы  $\Theta$  для темпоральных тематических моделей

$Y$  — моменты времени (например, годы публикаций),  
 $D_y \subset D$  — все документы, относящиеся к моменту  $y \in Y$ .  
 $n_y = \sum_{d \in D_y} n_d$  — доля коллекции, относящаяся к моменту  $y$ .

1. Разреживание  $p(t|y) = \sum_{d \in D_y} \theta_{td} \frac{n_d}{n_y}$  в каждый момент  $y$ :

$$R_1(\Theta) = \tau_1 \sum_{y \in Y} \text{KL}\left(\frac{1}{|T|} \parallel p(t|y)\right) \rightarrow \max.$$

2. Сглаживание  $p(y|t) = \sum_{d \in D_y} \theta_{td} \frac{n_d}{n_t}$  в соседние моменты  $y, y-1$ :

$$R_2(\Theta) = -\tau_2 \sum_{y \in Y} \sum_{t \in T} |p(y|t) - p(y-1|t)| \rightarrow \max.$$

## Время как модальность. Регуляризатор $\Phi$

**Проблема** регуляризатора  $\Theta$  в онлайнном EM-алгоритме:  
соседние по времени документы могут попасть в разные пакеты.

Документы содержат слова  $w \in W^1$  и время  $y \in W^2 = Y$   
 $W^2$  — модальность моментов времени (time stamps)

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

$$R_1(\Phi^2) = -\tau_1 \sum_{y \in Y} \sum_{t \in T} \ln \phi_{yt} \rightarrow \max.$$

2. Сглаживание  $p(y|t) = \phi_{yt}$  в соседние моменты  $y, y-1$ :

$$R_2(\Phi^2) = -\tau_2 \sum_{y \in Y} \sum_{t \in T} |\phi_{yt} - \phi_{y-1,t}| \rightarrow \max.$$

## 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.

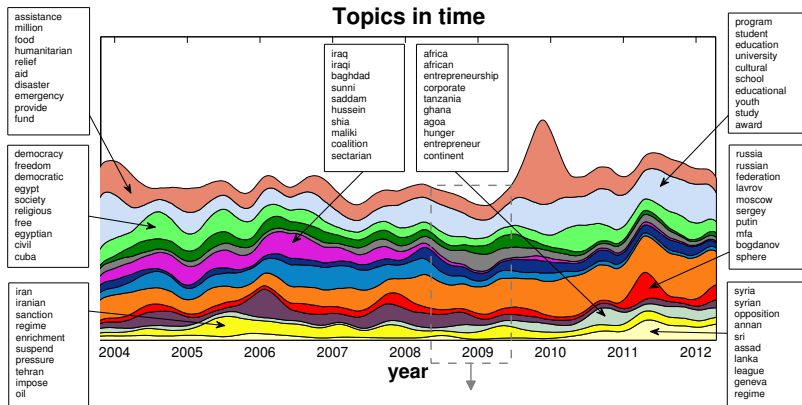
### Цели исследования:

- 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, моменты времени
- разреживание, сглаживание, декоррелирование
- сглаживание тем во времени

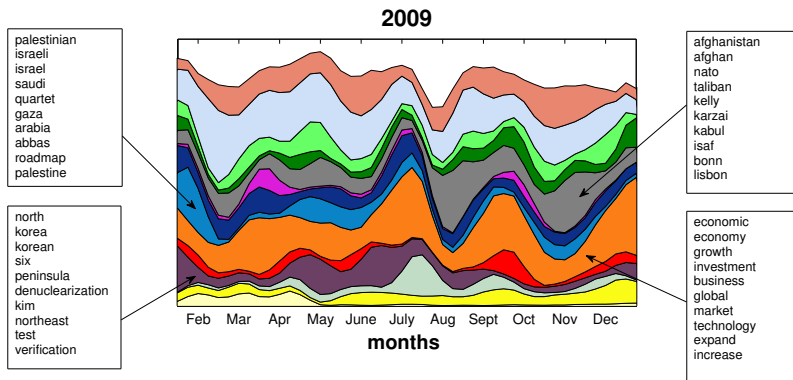
## Динамика тем во времени



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

## Динамика тем во времени

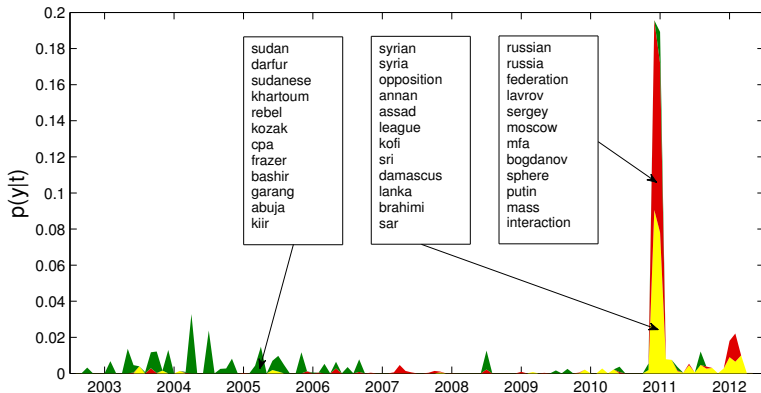
### Урупнение масштаба времени



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

## Динамика тем во времени

Пример: событийные темы и момент их совместного всплеска

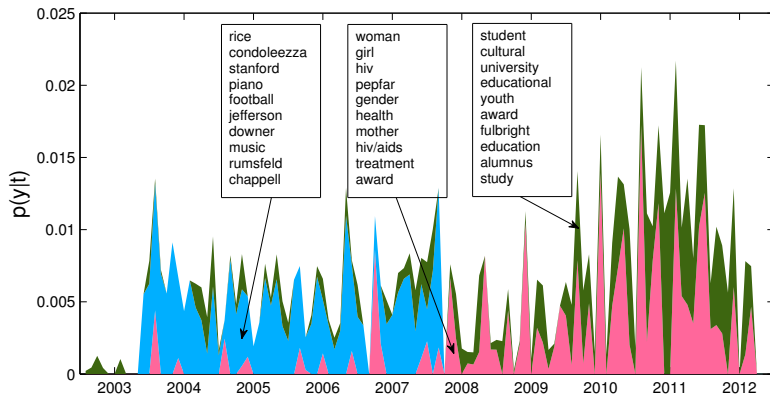


Никита Дойков. Адаптивная регуляризация вероятностных тематических моделей // ВКР бакалавра, 2015. ВМК МГУ.



## Динамика тем во времени

Примеры перманентных тем (сглаживание отключено)



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

## Гео-пространственные модели

Данные:  $\ell_d = (x_d, y_d)$  — геолокация (GPS) документа  $d$

Цели исследования:

- какие темы общие, какие специфичны для региона?
- какие есть похожие темы в других регионах?

Регуляризатор:

$$R(\Theta) = -\frac{T}{2} \sum_{(c,d)} w_{cd} \sum_{t \in T} (\theta_{td} - \theta_{tc})^2 \rightarrow \max,$$

$w_{cd}$  — вес пары  $(c, d)$ , близость геолокаций  $(x_c, y_c)$  и  $(x_d, y_d)$

---

*Zhijun Yin, Liangliang Cao, Jiawei Han, Chengxiang Zhai, Thomas Huang.*  
Geographical Topic Discovery and Comparison // WWW 2011.

## Пример: Food dataset

Где и что едят пользователи Flickr?



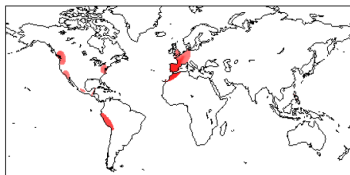
Chinese Food



Japanese Food



French Food



Spanish Food

---

*Zhijun Yin, Liangliang Cao, Jiawei Han, Chengxiang Zhai, Thomas Huang.*  
Geographical Topic Discovery and Comparison // WWW 2011.

## Выявление тематических сообществ

Граф  $\langle V, E \rangle$ , вершины  $v$  — подмножества  $D_v \subset D$ , например:

$D_v$  — отдельный документ  $v \equiv d$

$D_v$  — все статьи одного автора  $v$

$D_v$  — все посты из одного географического региона  $v$

Тематика вершины:

$$p(t|v) = \sum_{d \in D_v} p(t|d)p(d|v) = \sum_{d \in D_v} \theta_{td} \frac{n_d}{n_v}$$

Регуляризатор NetPLSA, при заданных весах  $w_{uv}$  рёбер  $(u, v)$ :

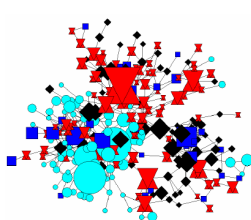
$$R(\Theta) = -\frac{\tau}{2} \sum_{(u,v) \in E} w_{uv} \sum_{t \in T} (p(t|v) - p(t|u))^2 \rightarrow \max$$

---

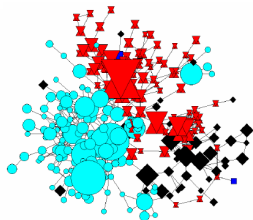
Qiaozhu Mei, Deng Cai, Duo Zhang, ChengXiang Zhai. Topic Modeling with Network Regularization // WWW-2008, Pp. 101–110.

## Примеры тематических сообществ

Регуляризация тем по соавторству в статьях конференций:

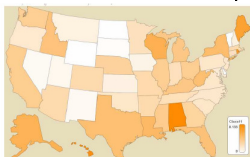


PLSA

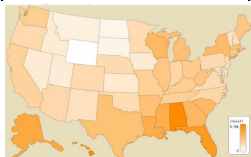


NetPLSA

Регуляризация тем по географической близости регионов:



With PLSA



With NetPLSA

## Направленные связи

**Проблема:** квадратичный регуляризатор NetPLSA игнорирует направленность связей  $u \rightarrow v$ .

**Предположение:** направление связи  $u \rightarrow v$  означает, что распределение  $p(t|v)$  «подчиняется» распределению  $p(t|u)$ .

**Модель iTopicModel.**

Вместо квадратичного регуляризатора минимизируется KL-дивергенция  $KL(p(t|v) \parallel p(t|u))$ :

$$R(\Theta) = \frac{\tau}{2} \sum_{(u,v) \in E} w_{uv} \sum_{t \in T} p(t|v) \ln p(t|u) \rightarrow \max$$

---

*Yizhou Sun, Jiawei Han, Jing Gao, Yintao Yu.* iTopicModel: Information Network-Integrated Topic Modeling // 2009.

## Создатель или распространитель контента?

$n_a$  — число сообщений пользователя  $a$

$r_a$  — число ретвитов пользователя  $a$

$r_{ab}$  — сколько раз  $b$  сделал ретвит сообщения пользователя  $a$

$\theta_{ta} = p(t|a)$  — тематика  $a$  как создателя контента

$\theta'_{ta} = p'(t|a)$  — тематика  $a$  как распространителя контента

### Предположения:

- если  $b$  ретвитит  $a$ , то тематики  $\theta_{ta}$  и  $\theta'_{tb}$  близки
- если  $c$  ретвитит  $a$  и  $b$ , то тематики  $\theta_{ta}$  и  $\theta_{tb}$  близки
- если  $a$  и  $b$  ретвоят  $c$ , то тематики  $\theta'_{ta}$  и  $\theta'_{tb}$  близки

---

*Wayne Xin Zhao, Jinpeng Wang, Yulan He, Jian-Yun Nie, Xiaoming Li.*

Originator or Propagator? Incorporating Social Role Theory into Topic Models for Twitter Content Analysis // CIKM 2013.

## Создатель или распространитель контента?

Меры близости пар пользователей  $a$  и  $b$ :

$\text{sim}_1(a, b) = \frac{r_{ab}}{n_a + r_b - r_{ab}}$  — как непосредственно взаимодействующих

$\text{sim}_2(a, b) = \frac{\sum_c r_{ac} r_{bc}}{(\sum_c r_{ac}^2)^{1/2} (\sum_c r_{bc}^2)^{1/2}}$  — как создателей контента

$\text{sim}_3(a, b) = \frac{\sum_c r_{ca} r_{cb}}{(\sum_c r_{ca}^2)^{1/2} (\sum_c r_{cb}^2)^{1/2}}$  — как распространителей контента

Регуляризаторы:

$$R_1(\Theta) = \tau_1 \sum_{(a,b) \in E} \text{sim}_1(a, b) \sum_{t \in T} (\theta_{ta} - \theta'_{tb})^2 \rightarrow \max;$$

$$R_2(\Theta) = \tau_2 \sum_{(a,b) \in E} \text{sim}_2(a, b) \sum_{t \in T} (\theta_{ta} - \theta_{tb})^2 \rightarrow \max;$$

$$R_3(\Theta) = \tau_3 \sum_{(a,b) \in E} \text{sim}_3(a, b) \sum_{t \in T} (\theta'_{ta} - \theta'_{tb})^2 \rightarrow \max;$$



- Регуляризаторы позволяют нацелить тематическую модель на классификацию, регрессию, выявление связей
- Разнообразные типы сущностей, встречающиеся в документах, удобно представлять модальностями
- Разнообразные тематические модели удобно строить, комбинируя регуляризаторы и модальности