Вероятностные тематические модели
Лекция 5. Модальности в ARTM

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 • 22 марта 2018

Содержание

1. **Мультиязычные тематические модели**
   - Параллельные и сравнимые тексты
   - Двуязычные словари
   - Кросс-язычный поиск
2. **Зависимости, корреляции, связи**
   - Классификация и регрессия
   - Модель CTM (Correlated Topic Model)
   - Гиперссылки, цитирование, влияние
3. **Социальные сети**
   - Тематические сообщества
   - Направленные связи
   - Социальные роли пользователей

Напоминание. Мультимодальная ARTM: постановка задачи

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 $\underset{W^m \times T}{\Phi^m}$ и $\underset{T \times D}{\Theta}$:
$\phi_{wt} = p(w|t)$ — вероятности токена w в каждой теме t,
$\theta_{td} = p(t|d)$ — вероятности тем t в каждом документе d.

Критерий максимума регуляризованного правдоподобия:

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

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

Vorontsov K.V., Frei O.I., Apishev M.A., Romov P.A., Suvorova M.A., Yanina A.O. Non-Bayesian Additive Regularization for Multimodal Topic Modeling of Large Collections // CIKM. Workshop on Topic Models, 2015.

Воронцов К.В. Вероятностное тематическое моделирование: обзор моделей. 2017. http://www.MachineLearning.ru/wiki/images/d/d5/Voron17survey-artm.pdf

Напоминание. Мультимодальная ARTM: EM-алгоритм

Максимизация log правдоподобия с регуляризатором R:

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

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

EM-алгоритм: метод простой итерации для системы уравнений

$$\text{E-шаг:}\quad p_{tdw} \equiv p(t|d,w) = \underset{t\in T}{\text{norm}}\left(\phi_{wt}\theta_{td}\right);$$

$$\text{M-шаг:}\quad \begin{cases} \phi_{wt} = \underset{w\in W^m}{\text{norm}}\left(n_{wt} + \phi_{wt}\dfrac{\partial R}{\partial \phi_{wt}}\right), \quad n_{wt} = \sum\limits_{d\in D}\tilde{n}_{dw}p_{tdw}; \\ \theta_{td} = \underset{t\in T}{\text{norm}}\left(n_{td} + \theta_{td}\dfrac{\partial R}{\partial \theta_{td}}\right), \quad n_{td} = \sum\limits_{w\in d}\tilde{n}_{dw}p_{tdw}; \end{cases}$$

где $\underset{t\in T}{\text{norm}}(x_t) = \dfrac{\max\{x_t,0\}}{\sum\limits_{s\in T}\max\{x_s,0\}}$ — операция нормировки вектора.

Напоминание. Онлайновый EM-алгоритм для ARTM

Коллекция D разбивается на пакеты D_b, $b = 1, \ldots, B$, которые могут обрабатываться параллельно и/или распределённо.

Вход: коллекция документов D,
параметры $\delta \equiv \texttt{decay\_weight}$, $\alpha \equiv \texttt{apply\_weight}$;
Выход: матрица Φ;

инициализировать ϕ_{wt} для всех $w \in W$, $t \in T$;
$n_{wt} := 0$, $\tilde{n}_{wt} := 0$ для всех $w \in W$, $t \in T$;
для всех пакетов D_b, $b = 1, \ldots, B$
  $(\tilde{n}_{wt}) := (\tilde{n}_{wt}) + \textbf{ProcessBatch}(D_b, \Phi)$;
  если пора обновить матрицу Φ **то**
    $n_{wt} := \delta n_{wt} + \alpha \tilde{n}_{wt}$ для всех $w \in W$, $t \in T$;
    $\phi_{wt} := \underset{w \in W^m}{\text{norm}} \left(n_{wt} + \phi_{wt} \frac{\partial R}{\partial \phi_{wt}} \right)$ для всех $m \in M$, $w \in W^m$, $t \in T$;
    $\tilde{n}_{wt} := 0$ для всех $w \in W$, $t \in T$;

Напоминание. Онлайновый EM-алгоритм для ARTM

Функция **ProcessBatch** обрабатывает пакет документов D_b, не меняя матрицу Φ, и выдаёт счётчики токенов в темах $\tilde{n}_{wt}$.

Вход: пакет документов D_b, матрица $\Phi = (\phi_{wt})$;
Выход: матрица счётчиков $(\tilde{n}_{wt})_{W \times T}$;

$\tilde{n}_{wt} := 0$ для всех $w \in W,\ t \in T$;
для всех $d \in D_b$
  инициализировать $\theta_{td} := \frac{1}{|T|}$ для всех $t \in T$;
  повторять
    $p_{tdw} := \underset{t \in T}{\text{norm}}\left(\phi_{wt}\theta_{td}\right)$ для всех $w \in d,\ t \in T$;
    пост-обработка матрицы $(p_{tdw})_{T \times n_d}$ при необходимости;
    $\theta_{td} := \underset{t \in T}{\text{norm}}\left(\sum_{w \in d} \tau_{m(w)} n_{dw} p_{tdw} + \theta_{td} \frac{\partial R}{\partial \theta_{td}}\right)$ для всех $t \in T$;
  пока θ_d не сойдётся;
  $\tilde{n}_{wt} := \tilde{n}_{wt} + \tau_{m(w)} n_{dw} p_{tdw}$ для всех $w \in W,\ t \in T$;

Параллельные и сравнимые корпуса текстов

Parallel — точный перевод (с выравниванием предложений), пример: EuroParl, протоколы европарламента, 21 язык.

Comparable — не перевод, а пересказ на другом языке, пример: Википедия.

W^ℓ — словарь языка ℓ из множества языков L.

Модель ML-P (MultiLingual Parallel)

- каждый язык — отдельная модальность
- $\theta_{td} = p(t|d)$ общее для всех связных документов $d = \bigsqcup_{\ell \in L} d^\ell$

Дополнительные данные — двуязычные словари:

- $\Pi_k(w) \subset W^k$ — все переводы слова $w \in W^\ell$ в языке k

---

I. Vulić, W. De Smet, J. Tang, M.-F. Moens. Probabilistic topic modeling in multilingual settings: an overview of its methodology and applications. 2015

Пример тем. Мультиязычная модель Википедии

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

Тема №68				Тема №79			
research	4.56	институт	6.03	goals	4.48	матч	6.02
technology	3.14	университет	3.35	league	3.99	игрок	5.56
engineering	2.63	программа	3.17	club	3.76	сборная	4.51
institute	2.37	учебный	2.75	season	3.49	фк	3.25
science	1.97	технический	2.70	scored	2.72	против	3.20
program	1.60	технология	2.30	cup	2.57	клуб	3.14
education	1.44	научный	1.76	goal	2.48	футболист	2.67
campus	1.43	исследование	1.67	apps	1.74	гол	2.65
management	1.38	наука	1.64	debut	1.69	забивать	2.53
programs	1.36	образование	1.47	match	1.67	команда	2.14

Асессор оценил 396 тем из 400 как хорошо интерпретируемые.

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

Пример тем. Мультиязычная модель Википедии

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

Тема №88				Тема №251			
opera	7.36	опера	7.82	windows	8.00	windows	6.05
conductor	1.69	оперный	3.13	microsoft	4.03	microsoft	3.76
orchestra	1.14	дирижер	2.82	server	2.93	версия	1.86
wagner	0.97	певец	1.65	software	1.38	приложение	1.86
soprano	0.78	певица	1.51	user	1.03	сервер	1.63
performance	0.78	театр	1.14	security	0.92	server	1.54
mozart	0.74	партия	1.05	mitchell	0.82	программный	1.08
sang	0.70	сопрано	0.97	oracle	0.82	пользователь	1.04
singing	0.69	вагнер	0.90	enterprise	0.78	обеспечение	1.02
operas	0.68	оркестр	0.82	users	0.78	система	0.96

Асессор оценил 396 тем из 400 как хорошо интерпретируемые.

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

Регуляризация по двуязычным словарям. Модель ML-TD

Гипотеза. Если $u \in \Pi_k(w)$, то тематика слов w и u близка:

$$\mathrm{KL}\big(\hat{p}(t|u) \,\big\|\, p(t|w)\big) \to \min,$$

где $\hat{p}(t|u) = \frac{n_{ut}}{n_u}$, $\quad p(t|w) = p(w|t)\frac{p(t)}{p(w)} = \phi_{wt}\frac{n_t}{n_w}$.

Модель ML-TD (MultiLingual Translation Dictionary)

$$R(\Phi) = \tau \sum_{\ell,k \in L} \sum_{w \in W^\ell} \sum_{u \in \Pi_k(w)} \sum_{t \in T} n_{ut} \ln \phi_{wt} \;\to\; \max_{\Phi}.$$

Недостатки. Модель ML-TD не учитывает два обстоятельства:

- тематику омонимов сближать не нужно,
- слово может иметь разные переводы в разных темах.

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

Матрица вероятностей переводов. Модель ML-TDP

Гипотеза. Переводы слов зависят от тем: $\pi_{uwt}^{k\ell} = p(u|w,t)$, темы согласуются в разных языках через переводы слов:

$$\mathrm{KL}\big(\hat{p}(u|t) \,\big\|\, p(u|t)\big) \to \min;$$

$\hat{p}(u|t) = \frac{n_{ut}}{n_t}$ — частотная оценка по модальности (языку) k,
$p(u|t)$ — модель темы t в языке k по языку ℓ:

$$p(u|t) = \sum_{w \in \Pi_\ell(u)} p(u|w,t)p(w|t) = \sum_{w \in \Pi_\ell(u)} \pi_{uwt}^{k\ell}\phi_{wt}.$$

Модель ML-TDP (MultiLingual Translation Dictionary Probability)

$$R(\Phi,\Pi) = \tau \sum_{\ell,k \in L} \sum_{u \in W^k} \sum_{t \in T} n_{ut} \ln \sum_{w \in \Pi_\ell(u)} \pi_{uwt}^{k\ell}\phi_{wt} \to \max_{\Phi,\Pi}.$$

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

Формулы M-шага для моделей ML-TD и ML-TDP

ML-TD (MultiLingual Translation Dictionary):

$$\phi_{wt} = \underset{w \in W^\ell}{\text{norm}} \left(n_{wt} + \tau \sum_{k \in L \setminus \ell} \sum_{u \in \Pi_k(w)} n_{ut} \right)$$

ML-TDP (MultiLingual Translation Dictionary Probability):

$$\phi_{wt} = \underset{w \in W^\ell}{\text{norm}} \left(n_{wt} + \tau \sum_{k \in L \setminus \ell} \sum_{u \in \Pi_k(w)} \pi^{\ell k}_{wut} n_{ut} \right)$$

$$\pi^{k\ell}_{uwt} = \underset{u \in W^k}{\text{norm}} \left(\pi^{\ell k}_{wut} n_{ut} \right)$$

Смысл регуляризации:

условные вероятности $\phi_{wt} = p(w|t)$ согласуются с их частотными оценками по словам других языков

Тематические переводы слов $\pi^{k\ell}_{uwt} = p(u|w,t)$

Темы, в которых $p(\text{«sum»}|\text{«сумма»},t) > 0.9$

Тема №6		Тема №12		Тема №20	
множество	set	математика	triangle	вектор	vector
пространство	space	треугольник	square	координата	coordinate
группа	point	теорема	number	пространство	field
точка	left	точка	point	преобразование	tensor
элемент	limit	математический	theorem	базис	transform
функция	symmetry	угол	angle	тензор	basis
предел	function	координата	mathematics	сила	space
отображение	open	экономика	real	векторный	force
симметрия	property	число	theory	точка	rotation
открытый	topology	квадрат	geometry	система	thermometer

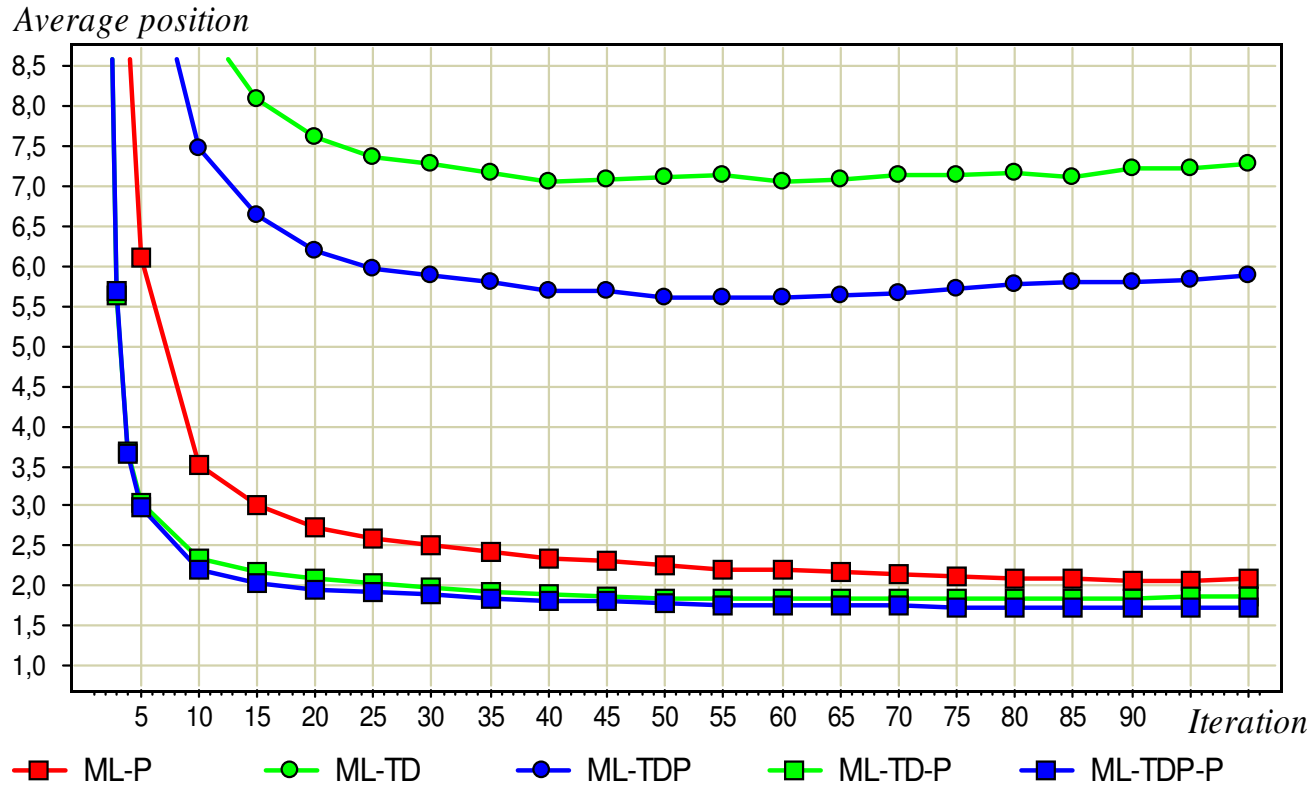
Темы, в которых $p(\text{«total»}|\text{«сумма»},t) > 0.9$

Тема №5		Тема №19		Тема №22	
орбита	space	программный	software	игра	game
аппарат	nasum	версия	version	видеосигнал	character
космический	orbit	работа	news	игрок	video
земля	instrument	компания	company	фильм	player
поверхность	earth	анонимный	work	головоломка	series
солнечный	surface	примечание	note	серия	puzzle
станция	solar	терминатор	release	качество	movie
запуск	system	журнал	support	шахматы	jason
система	landing	рей	terminator	джейсон	world
атмосфера	camera	персонаж	anonymous	буква	chess

Кросс-язычный поиск: ищем документ по его переводу

Wiki: $|D| = 586$, категория «Математика», $|T| = 100$,
$|W^{\text{рус}}| = 19\,305$, $|W^{\text{eng}}| = 23\,413$, переводов 82 642 пар.

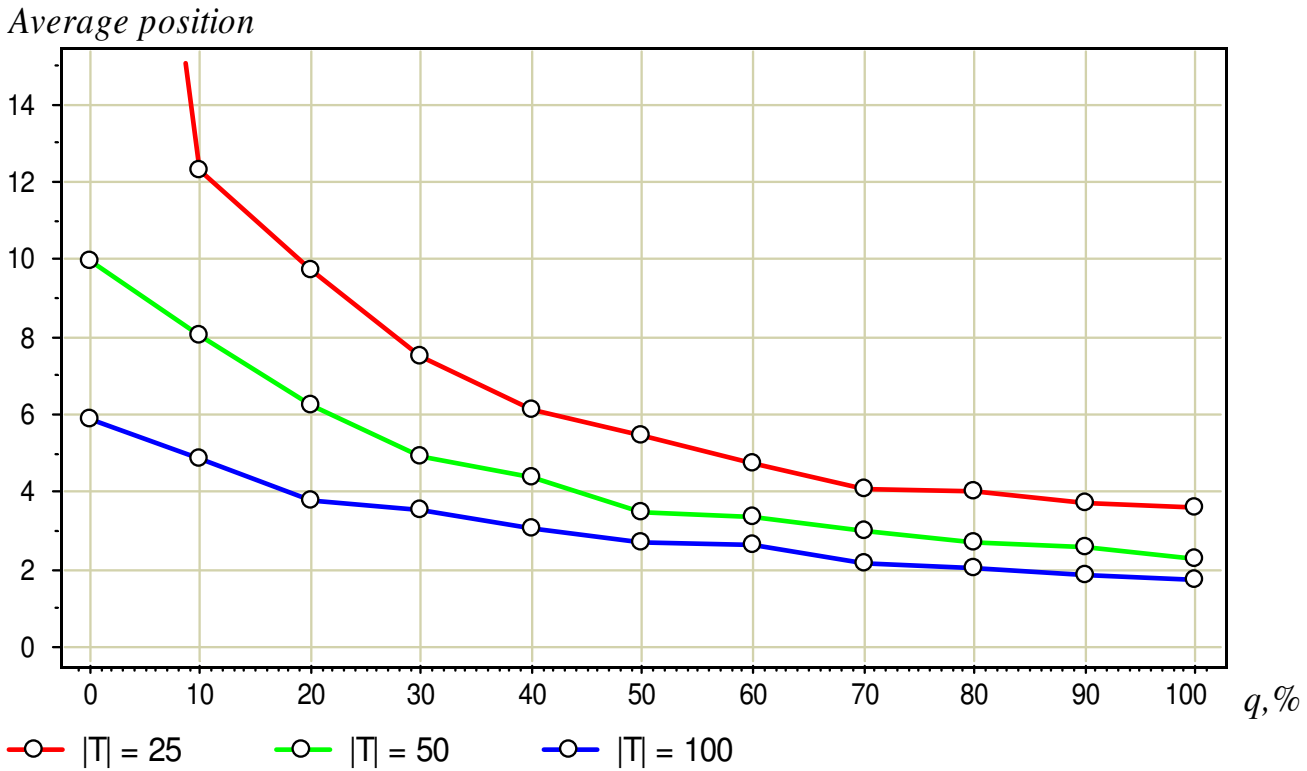
Качество поиска — средняя позиция перевода в выдаче:



Кросс-язычный поиск: ищем документ по его переводу

Wiki: $|D| = 586$, категория «Математика», $|T| = 25, 50, 100$,
$|W^{\text{рус}}| = 19\,305$, $|W^{\text{eng}}| = 23\,413$, переводов 82 642 пар.

Зависимость средней позиции перевода в выдаче
от числа тем $|T|$ и доли q параллельных текстов в коллекции:



Резюме по мультиязычным моделям

- Главное чудо: для построения мультиязычных тем достаточно иметь сравнимые корпуса.
- Сравнимая коллекция является более сильным источником многоязычной информации, чем словарь переводов (!)
- Модель с вероятностями переводов — самая сильная
- Не обязательно, чтобы все документы имели параллельные
- Главное применение — по запросу на одном языке ищем:
  — тексты на другом языке — *кросс-язычный поиск*,
  — тексты на всех языках — *мульти-язычный поиск*.
- Применение в статистическом машинном переводе:
  выбор варианта перевода согласно тематике документа.

Тематическая модель классификации (категоризации)

Обучающие данные: C — множество классов (категорий);
$C_d \subseteq C$ — классы, к которым d относится;
$C'_d \subseteq C$ — классы, к которым d не относится.

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

Правдоподобие вероятностной модели бинарных данных:

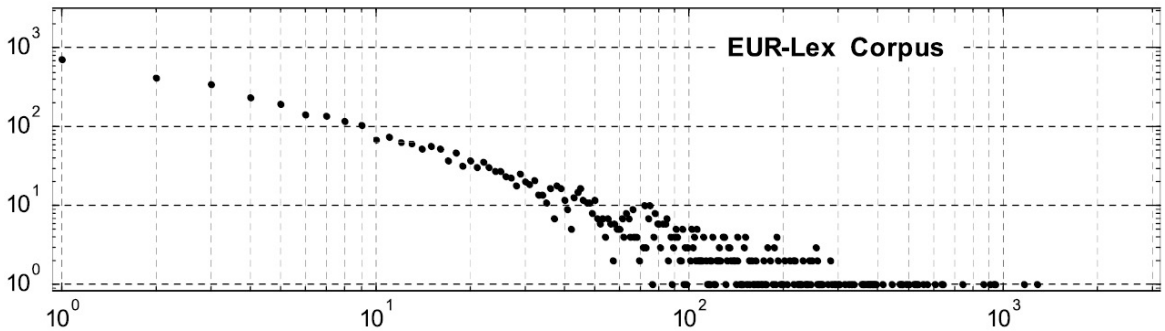
$$R(\Phi, \Theta) = \tau \sum_{d \in D} \sum_{c \in C_d} \ln \sum_{t \in T} \phi_{ct}\theta_{td} + \\ + \tau \sum_{d \in D} \sum_{c \in C'_d} \ln\Big(1 - \sum_{t \in T} \phi_{ct}\theta_{td}\Big) \;\to\; \max$$

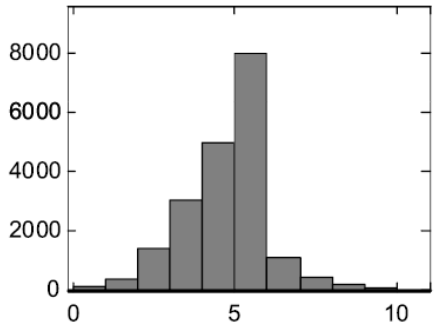
При $C'_d = \varnothing$, $n_{dc} = [c \in C_d]$ это правдоподобие модальности C.

Rubin T. N., Chambers A., Smyth P., Steyvers M. Statistical topic models for multi-label document classification // Machine Learning, 2012, no. 88 (1–2).

Эксперимент. Категоризация коллекции EUR-Lex

- EUR-Lex: $|D| = 19\,800$ документов — законы Евросоюза
- Две модальности: W^1 слова (21K), 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, no. 88 (1–2).

Эксперимент. Категоризация коллекции EUR-Lex

Регуляризаторы:

- Равномерное сглаживание Θ
- Равномерное сглаживание матрицы слова–темы Φ^1
- *Label regularization* для матрицы категории–темы Φ^2:

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

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] — среди байесовских моделей ближайший аналог ARTM для классификации

Критерии качества [Rubin 2012]:

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

Результаты сравнения:

	AUC-PR⇑	AUC⇑	OneErr⇓	IsErr⇓
BigARTM	**52.9**	98.0	**27.1**	**94.2**
DLDA [Rubin 2012]	49.2	**98.2**	32.0	97.2
SVM	43.5	97.5	31.6	98.1

Тематическая модель регрессии

Обучающие данные: $y_d \in \mathbb{R}$ для всех документов $d \in D$.

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

Регуляризатор — среднеквадратичная ошибка (МНК):

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

Подставляем, получаем формулы M-шага:

$$\theta_{td} = \operatorname*{norm}_t \left(n_{td} + \tau v_t \theta_{td} \left(y_d - \sum_{s \in T} v_s \theta_{sd} \right) \right);$$

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

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

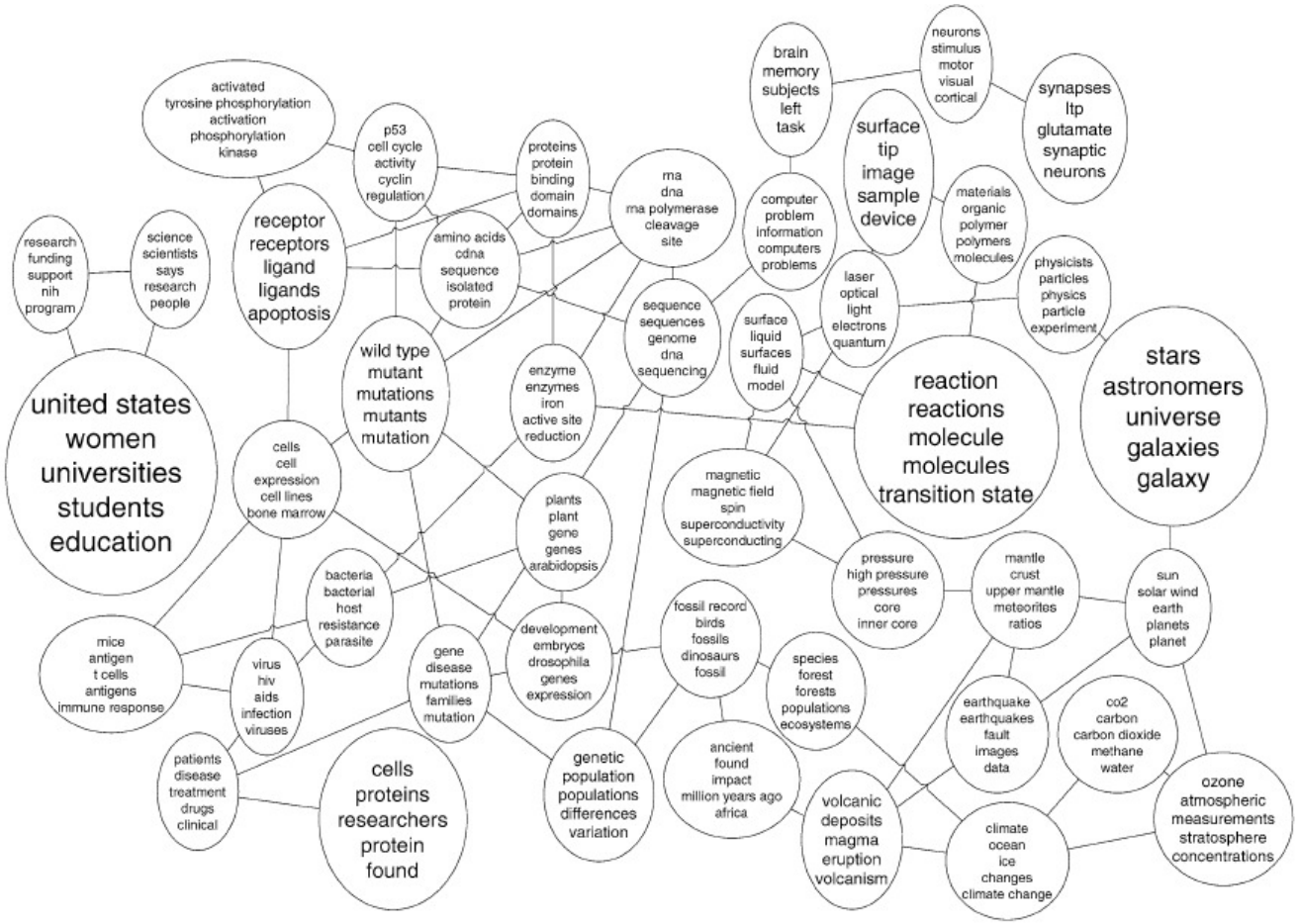
Прогнозирование скачков цен на финансовых рынках

d — текст новости

y_d — изменение цены в последующие 10–60 минут

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

CTM: модель коррелированных тем



David Blei, John Lafferty. A Correlated Topic Model of SCIENCE // Annals of Applied Statistics, 2007. Vol. 1, Pp. 17-35.

Многомерное лог-нормальное распределение

Мотивация. Темы могут коррелировать: «статьи по археологии чаще связаны с историей и геологией, чем с генетикой».

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

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

где $\eta_d = (\eta_{td})_{t\in T}$, $\eta_{td} = \ln \theta_{td} + C_d$ — векторы документов, определённые с точностью до константы C_d, не зависящей от t,

μ, S — параметры гауссовского распределения.

Регуляризатор модели коррелированных тем CTM

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

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

Регуляризатор с параметрами μ, S:

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

Формулы M-шага (S, μ можно обновлять намного реже, чем Θ):

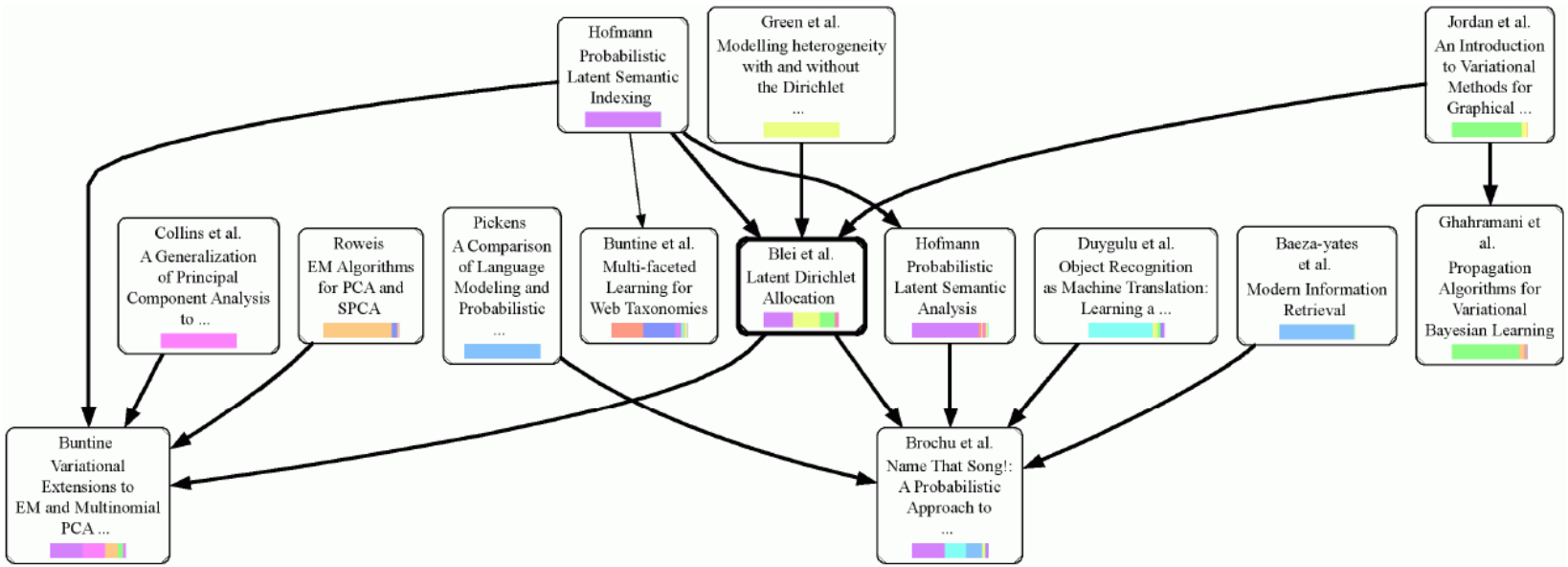
$$\theta_{td} = \operatorname*{norm}_{t \in T} \Big(n_{td} - \tau \sum_{s \in T} S^{-1}_{ts} \big(\ln \theta_{sd} - \mu_s \big) \Big);$$

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

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

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 Θ 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 θ_d, θ_c:

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

Подставляем, получаем ещё один вариант сглаживания:

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

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

Связи как модальность. Регуляризатор Ф

Проблема учёта связей в онлайновом 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:

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

Другой вариант — сумма ковариационных регуляризаторов:

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

Задача выявления 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:

$$p(t|v) = \sum_{d \in D_v} p(t|d)p(d|v) = \frac{1}{|D_v|} \sum_{d \in D_v} \theta_{td}$$

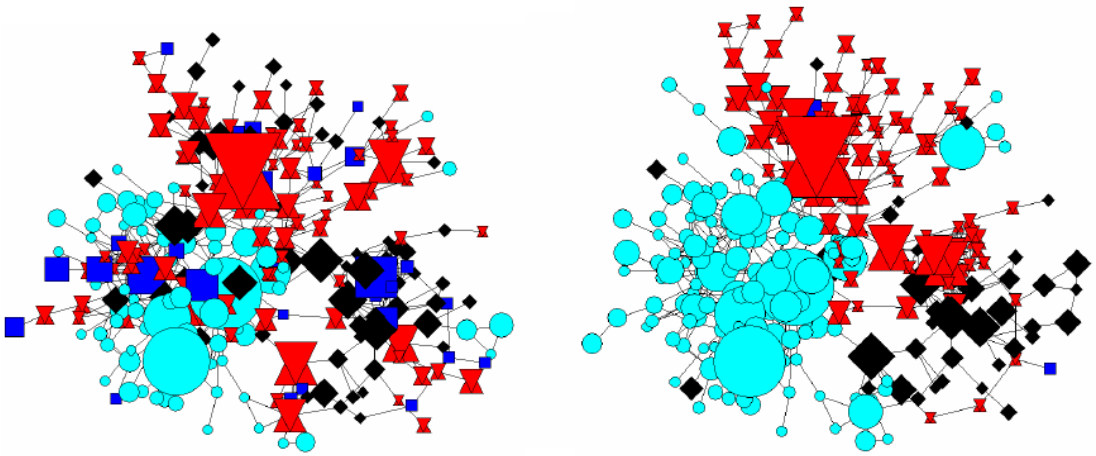
Регуляризатор NetPLSA, при заданных весах w_{uv} рёбер (u, v):

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

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

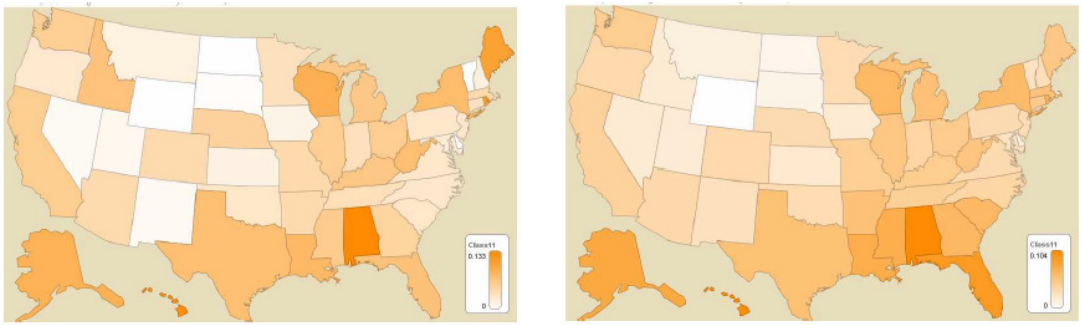
Примеры тематических сообществ

D_v — все статьи автора v на четырёх конференциях:



PLSA NetPLSA

D_v — все посты про ураган Катрина из штата v:



With PLSA With NetPLSA

От NetPLSA к модальности вершин графа

Проблема регуляризатора Θ в онлайновом EM-алгоритме: связанные документы могут попасть в разные пакеты.

$W^2 = V$ — модальность вершин графа $\langle V, E \rangle$.
В каждый документ $d \in D_v$ добавляется токен v.

Тематика вершины:

$$p(t|v) = p(v|t)\frac{p(t)}{p(v)} = \phi_{vt}\frac{n_t}{n_v}$$

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

$$R(\Phi^2) = -\frac{\tau}{2}\sum_{(u,v)\in E} w_{uv}\sum_{t\in T} n_t^2\left(\frac{\phi_{vt}}{n_v} - \frac{\phi_{ut}}{n_u}\right)^2 \to \max_{\Phi}$$

Виктор Булатов. Использование графовой структуры в тематическом моделировании // Магистерская диссертация, ФИВТ МФТИ, 2016.

Направленные связи

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

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

Модель iTopicModel. В отличие от NetPLSA, минимизируется не квадратичный критерий, а дивергенция $\mathrm{KL}\big(p(t|v) \,\big\|\, p(t|u)\big)$:

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

причём $p(t|v)$ можно выразить и через Θ, и через Φ^2.

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

Создатель или распространитель контента?

Документ $a \in D$ — все твиты, созданные пользователем a
Документ $b \in D$ — все ретвиты пользователя b
n_a — число сообщений пользователя a
r_b — число ретвитов пользователя b
r_{ab} — сколько раз b сделал ретвит сообщения пользователя a
$\theta_{ta} = p(t|a)$ — тематика a в роли создателя контента
$\theta'_{tb} = p'(t|b)$ — тематика b в роли распространителя контента

Предположения:

- если b ретвитит a, то тематики θ_{ta} и θ'_{tb} близки
- если c ретвитит a и b, то тематики θ_{ta} и θ_{tb} близки
- если a и b ретвитят c, то тематики θ'_{ta} и θ'_{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

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

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

Регуляризаторы:

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

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

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

Переход к модальностям создателей и распространителей

Проблема регуляризатора Θ в онлайновом EM-алгоритме: связанные пользователи могут попасть в разные пакеты.

Документ $d \in D$ — отдельный твит, содержащий:
$a_d \in A$ — один токен модальности Φ^A создателя,
$b \in B_d \subset B$ — токены модальности Φ^B распространителей,
$A \equiv B$ — множество всех пользователей социальной сети.

Регуляризаторы над $p(t|a) = \phi^A_{at} \frac{n_a}{n_t}$ и $p(t|b) = \phi^B_{bt} \frac{n_b}{n_t}$:

$$R_1(\Phi) = \tau_1 \sum_{(a,b)} \text{sim}_1(a,b) \sum_{t \in T} \left(\phi^A_{at} \frac{n_a}{n_t} - \phi^B_{bt} \frac{n_b}{n_t} \right)^2 \to \max;$$

$$R_2(\Phi) = \tau_2 \sum_{(a,b)} \text{sim}_2(a,b) \sum_{t \in T} \left(\phi^A_{at} \frac{n_a}{n_t} - \phi^A_{bt} \frac{n_b}{n_t} \right)^2 \to \max;$$

$$R_3(\Phi) = \tau_3 \sum_{(a,b)} \text{sim}_3(a,b) \sum_{t \in T} \left(\phi^B_{at} \frac{n_a}{n_t} - \phi^B_{bt} \frac{n_b}{n_t} \right)^2 \to \max;$$

Резюме

- Регуляризаторы позволяют нацелить тематическую модель на классификацию, регрессию, выявление связей
- Разнообразные типы сущностей, встречающиеся в документах, удобно представлять модальностями
- Регуляризаторы Θ, не удобные в онлайновом алгоритме, можно превращать в регуляризаторы Φ
- Разнообразные тематические модели удобно строить, комбинируя регуляризаторы и модальности