Построение иерархической модели крупной конференции

Александр Сергеевич Златов

Московский физико-технический институт
Факультет управления и прикладной математики
Кафедра интеллектуальных систем

Научный руководитель: д.ф.-м.н. В. В. Стрижов
Консультант: А. А. Кузьмин

15.06.2016.г

Цель исследования

Цель

- Построение тематической модели конференции и верификация экспертной модели.

Задача

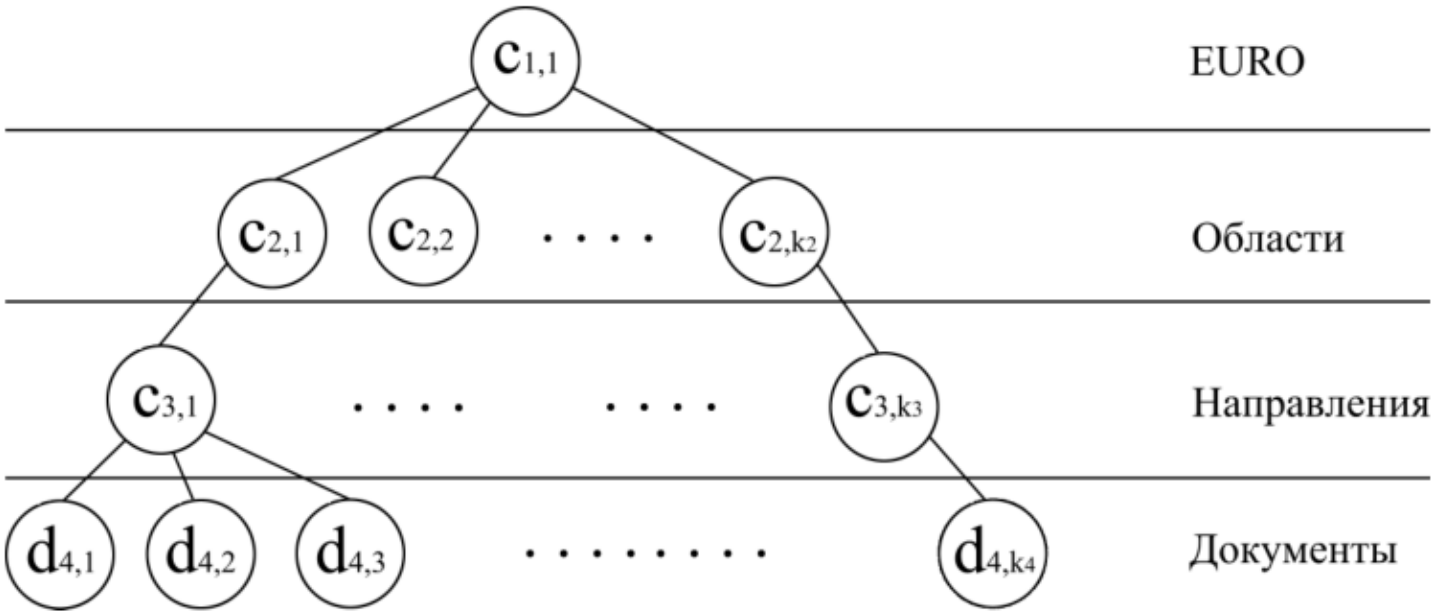
- Построить иерархическую тематическую модель крупной конференции по экспертной модели.

Метод

- Существующий метод DPM адаптируется для плоской кластеризации тезисов крупной конференции.
- На основании плоского метода строится дивизимный иерархический алгоритм кластеризации.

Структура крупной конференции

Структура крупной конференции на примере EURO



Литература

1. Ee-Peng Lim, Arindam Banerjee, Qi He, Kuiyu Chang. Keep it simple with time: A re-examination of probabilistic topic detection models. 2009.
2. Tu Z. Probabilistic boosting-tree: Learning discriminative models for classification, recognition, and clustering // Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on. – IEEE, 2005.
3. Kuzmin A.A., Aduenko A.A., Strijov V.V. Thematic Classification for EURO/IFORS Conference Using Expert Model // Conference of the International Federation of Operational Research Societies, 2014.

Постановка задачи

Матрица $\mathbf{X}$ содержит описания документов d_s из коллекции D. Слово x_j из словаря W встретилось в документе d_s k раз: $x_{s,j} = k,\ k \geq 0$

$$\mathbf{X} = \begin{pmatrix} x_{1,1} & \dots & x_{1,n} \\ \dots & \dots & \dots \\ x_{|D|,1} & \dots & x_{|D|,n} \end{pmatrix}.$$

Процент правильно классифицированных документов на уровне ℓ: $S_\ell = \frac{1}{k_\ell} \sum_{i=1}^{k_\ell} [c_{\ell i} = \tilde{c_{\ell i}}]$. Ставится задача $S \to \max$.

Дополнительные критерии качества

- Качество по релевантности экспертного кластера:
  $Q = \frac{1}{|D|} \sum_{j=1}^{|D|} \text{pos}(R(\mathbf{x}_j), \tilde{c_{\ell i}})$.
- Площадь под кривой AUC:
  $AUC = \frac{1}{k_\ell |D|} \sum^{k_\ell} \#\text{pos}(R(\mathbf{x}_j), \tilde{c_{\ell i}}) \leq i$.

Основные типы алгоритмов текстовой кластеризации

Тип моделей	Документ	Тема	Пример алгоритма
Жесткие	вектор	вектор	k-means
Описательно-вероятностные	вектор	вероятность	DPM
Смеси	вектор	распределение	mixture of Gaussian, vMF
Вероятностные	распределение	распределение	LDA

Разделяющая вероятностная модель (DPM)

Пусть W - весь словарь, а F - словарь информативных слов.
Пусть для кластера $c_{\ell i}$, документа $\mathbf{x}$ и слова x:
$p(c_{\ell i}, \mathbf{x}|x) = p(c_{\ell i}|x)p(\mathbf{x}|x)$, т.е. $p(c_{\ell i}, x|\mathbf{x}) = p(c_{\ell i}|x)$.

Тогда искомая вероятность

$$P(c_{\ell i}|\mathbf{x}) = \sum_{x \in W} P(c_{\ell i}, x|\mathbf{x})P(x|\mathbf{x}) =$$

$$\sum_{x \in F} P(c_{\ell i}, x|\mathbf{x})P(x|\mathbf{x}) + \sum_{x \in W \setminus F} P(c_{\ell i}, x|\mathbf{x})P(x|\mathbf{x}) =$$

$$\sum_{x \in F} P(c_{\ell i}|x)P(x|\mathbf{x}) + P(c_{\ell i}|\mathbf{x}) \sum_{x \in W \setminus F} P(x|\mathbf{x}).$$

Положим для каждого документа $\sum_{x \in W \setminus F} P(x|\mathbf{x}) = R$, тогда

$$P(c_{\ell i}|\mathbf{x}) = \frac{1}{1-R} \sum_{x \in F} P(c_{\ell i}|x)P(x|\mathbf{x}).$$

Алгоритм DPM

Полученное выражение можно переписать в виде:

$$P(c_{\ell i}|\mathbf{x}) = \frac{1}{1-R}\sum_{x\in F}\frac{P(x|c_{\ell i})P(c_{\ell i})}{\sum\limits_{c_{\ell j}\in C}P(x|c_{\ell j})P(c_{\ell j})}P(x|\mathbf{x}).$$

Оценки для величин в данном выражении: $P(x|\mathbf{x}) = \frac{x}{\|\mathbf{x}\|}$,
$P(c_{\ell j}) = \frac{N(c_{\ell j})}{N}$, $P(x|c_{\ell j}) = \frac{1}{N(c_{\ell j})}\sum\limits_{\mathbf{x}'\in c_{\ell i}}P(x|\mathbf{x}')$.

Введем обозначения $\mathrm{TF}'(w_i, d) = \frac{x_i}{\|\mathbf{x}\|}$, $\mathrm{IDF}'(w_i) = \sqrt{\frac{N}{\sum\limits_{\mathbf{x}\in D}\frac{x_i}{\|\mathbf{x}\|}}}$
и перейдём к представлению документа $x_i = \mathrm{TF}'(x_i, d)\cdot\mathrm{IDF}'(x_i)$.
Документ $\mathbf{x}$ принадлежит кластеру $c_{\ell i}$ с вероятностью

$$P(c_{\ell i}|\mathbf{x}) = \frac{1}{1-R}\cdot\frac{N(c_{\ell i})}{N}\mathbf{x}^{\mathrm{T}}\mathbf{c}_{\ell i}, \qquad \mathbf{c}_{\ell i} = \frac{1}{N(c_{\ell i})}\cdot\sum_{\mathbf{x}'\in c_{\ell i}}\mathbf{x}'.$$

Алгоритм DPM

Шаг 1. Пересчитываем центры кластеров

$$\mathbf{c}_{\ell i} = \frac{1}{N(c_{\ell i})} \cdot \sum_{\mathbf{x}' \in c_{\ell i}} \mathbf{x}'.$$

Шаг 2. Рассчитываем новые вероятности $P^{\text{new}}(c_{\ell i}|\mathbf{x})$ по старым $P(c_{\ell i}|\mathbf{x})$ и $\mathbf{c}_{\ell i}$

$$P^{\text{new}}(c_{\ell i}|\mathbf{x}) = \frac{1}{1-R} \cdot \frac{N(c_{\ell i})}{N} \mathbf{x}^{\mathrm{T}} \mathbf{c}_{\ell i}, \qquad \text{где} \qquad N(c_{\ell i}) = \sum_{x \in c_{\ell i}} P(c_{\ell i}|\mathbf{x}).$$

Шаг 3. Присваиваем вероятности для документов из обучающей выборки по правилу

$$P(c_{\ell i}|\mathbf{x}) = \begin{cases} 1 & \text{при } c_{\ell i} = \tilde{c_{\ell i}}, \\ 0 & \text{при } c_{\ell i} \neq \tilde{c_{\ell i}}. \end{cases}$$

Модификации алгоритма DPM

lDPM и nDPM. Данные алгоритмы повзоляют снизить влияние величины кластера в формуле для $P(c_{\ell i}|\mathbf{x})$.
На Шаге 2 алгоритма lDPM используется формула

$$P^{\text{new}}(c_{\ell i}|\mathbf{x}) = \frac{1}{1-R} \cdot \frac{\ln N(c_{\ell i})}{N} \mathbf{x}^{\mathrm{T}} \mathbf{c}_{\ell i}.$$

На Шаге 2 алгоритма nDPM используется формула

$$P^{\text{new}}(c_{\ell i}|\mathbf{x}) = \frac{1}{1-R} \cdot \frac{1}{N} \mathbf{x}^{\mathrm{T}} \mathbf{c}_{\ell i}.$$

hDPM. Для иерархической кластеризации алгоритм DPM запускается для второго уровня иерархии($\ell = 2$). После чего для каждого кластера запускается аналогичный алгоритм DPM, разделяющий документы уже по кластерам третьего уровня и так далее.

Построение модели конференции EURO

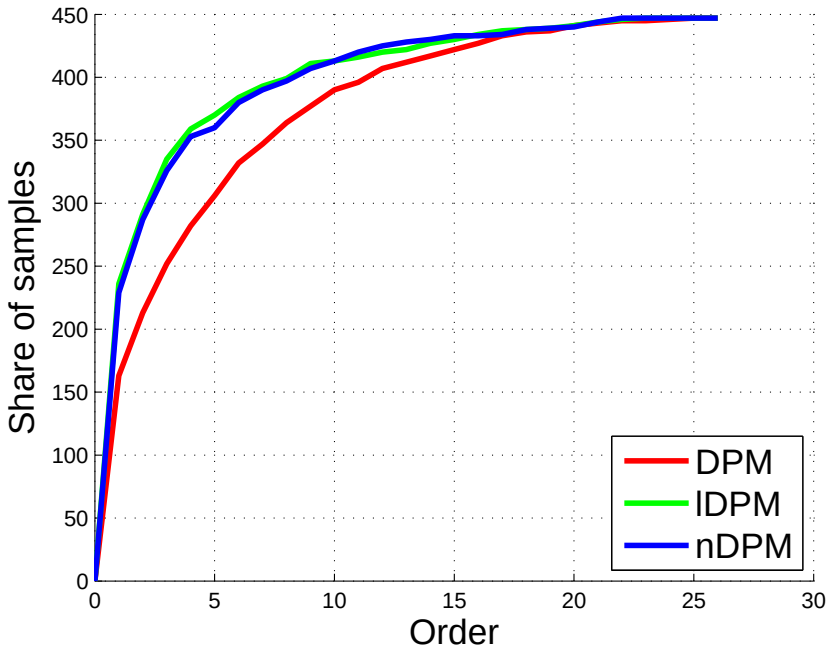
Сравнивалась работа алгоритмов DPM, lDPM и nDPM на наборе тезисов конференции EURO. Выборка состоит из 1342 тезисов. Объем словаря $n = 3479$ слов. Структура конференции: 26 областей и 114 направлений.

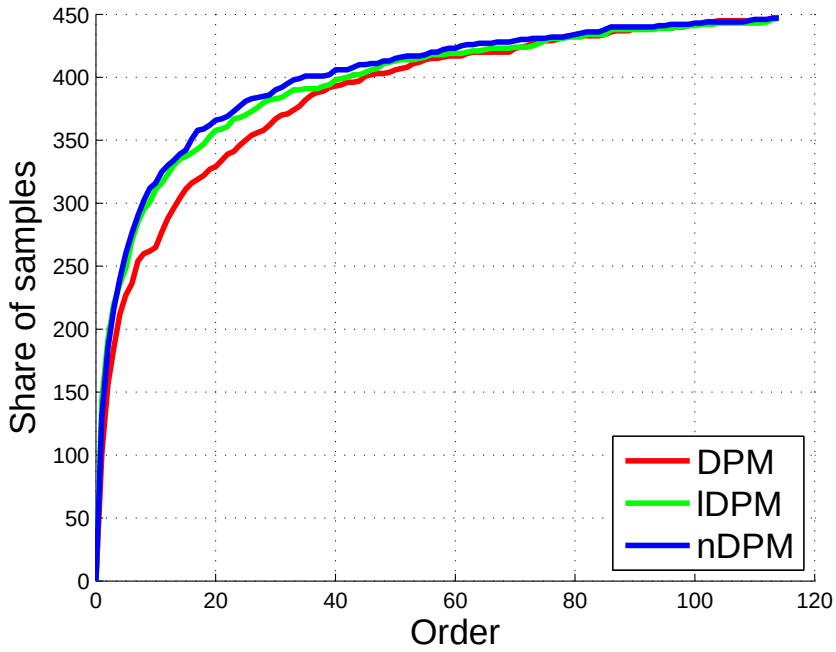
	DPM		lDPM		nDPM	
	область	напр.	область	напр.	область	напр.
S, %	35,6±2,1	24,0±1,8	**53,1±2,1**	**33,8±1,6**	51,9±1,8	30,6±1,8
Q	4,79±0,27	16,17±1,25	**3,35±0,16**	13,89±1,02	3,42±0,18	**12,79±1,02**
AUC	0,85±0,04	0,87±0,05	**0,91±0,05**	0,89±0,05	0,91±0,05	**0,90±0,05**

Иерархический выриант lDPM позволил улучшить плоский алгоритм.

	lDPM	hDPM
S, %	33,8±1,6	**35,2±1,6**

Качество модели





Верхние огибающие AUC.

Классификация сайтов индустриальной промышленности

Сравнивалась работа алгоритмов DPM, lDPM и nDPM на наборе сайтов индустриальной промышленности. Выборка состояла из 1076 тезисов. Объем словаря $n = 20278$ слов. Структура: 11 областей и 77 направлений.

	DPM		lDPM		nDPM	
	область	напр.	область	напр.	область	напр.
S, %	29.25±1,48	24.21±1,27	**65.41±2,36**	**49.06±1,87**	64.78±2,34	48.11±1,62

Иерархический выриант lDPM на этих данных работает хуже, чем плоский алгоритм.

	lDPM	hDPM
S, %	**49.06±2,1**	48.70±2,0

Заключение и публикации

Заключение

- Построена модель конференции по адаптированным алгоритмам DPM.
- Снижение влияния величины кластера позволило повысить качество построения модели по сравнению с оригинальным DPM.
- Дивизимный иерархический алгоритм на основании плоского алгоритма увеличил качество построенной модели.

Публикации

- Златов А. С., Кузьмин А. А. // Построение иерархической тематической модели крупной конференции // Искусственный интеллект и принятие решений, 2016. (Подана в журнал)