Построение иерархических тематических моделей крупных конференций

А. А. Кузьмин, В. В. Стрижов

Московский физико-технический институт

Москва, октябрь 2016 г.

Задача построения тематической модели

Задано

- Коллекция документов D
- Иерархическая структура коллекции $\mathcal{C}$
- Экспертная классификация для подмножества документов $\mathbf{Z}$

Определение

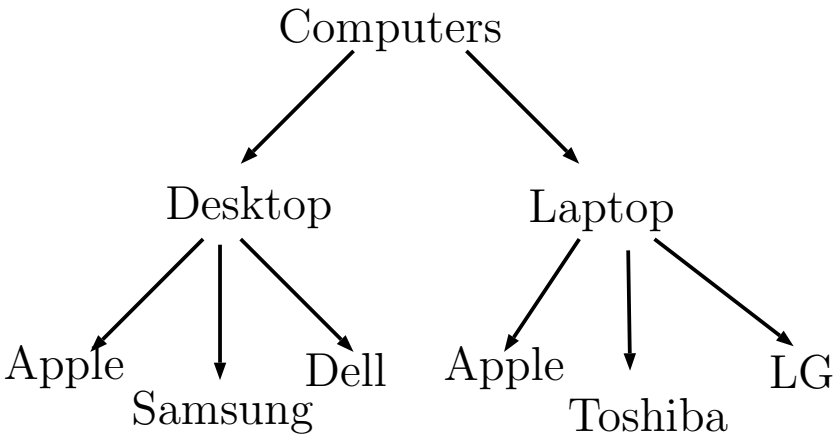
Тематической моделью M коллекции D со структурой $\mathcal{C}$ называется отображение множества документов D в множество кластеров $\mathcal{C}$.

Задача

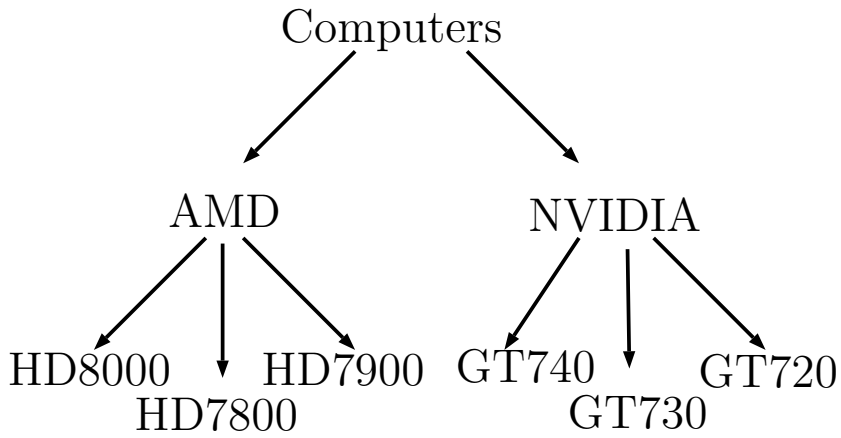
Построить тематическую модель M, согласованную с экспертной кластеризацией $\mathbf{Z}$ подмножества D.

Согласованность с экспертной моделью

D – коллекция статей из журнала о компьютерах.



Оптимальная модель.



Модель, необходимая эксперту.

Цель:
Выявить признаки, по которым эксперт разделил часть коллекции на классы и используя данные признаки классифицировать оставшиеся документы.

Построение списка индексов кластеров, ранжированного по релевантности документу

Определение

Пусть $q(\mathbf{x}) \in S^{k_h}$ – перестановка, соответствующая сортировке кластеров нижнего уровня h в порядке убывания релевантности документу $\mathbf{x}$, где k_h – количество кластеров.

Пример: $q(\mathbf{x}) = \{3, 1, \ldots, 6\}$.

Определение

Пусть $R : \mathbb{R}^n \to S^{k_h}$ – оператор релевантности, ставящий в соответствие каждому документу $\mathbf{x} \in \mathbb{R}^n$ перестановку $q(\mathbf{x}) \in S^{k_h}$.

Определение

Пусть $\mathrm{pos}(q, j) : S^q \times \{1, 2, \ldots, q\} \to \{1, 2, \ldots, q\}$ – функция позиции, возвращающая индекс числа j в перестановке q.

Пример: $\mathrm{pos}(q, 1) = 2$.

Критерии качества $Q(R)$ и $\mathrm{AUC\ CH}(R)$

Критерий качества $Q(R)$

Пусть $Q(R)$ – усредненная позиция экспертного кластера $z_{j,h}$ в перестановке $R(\mathbf{x}_j)$:

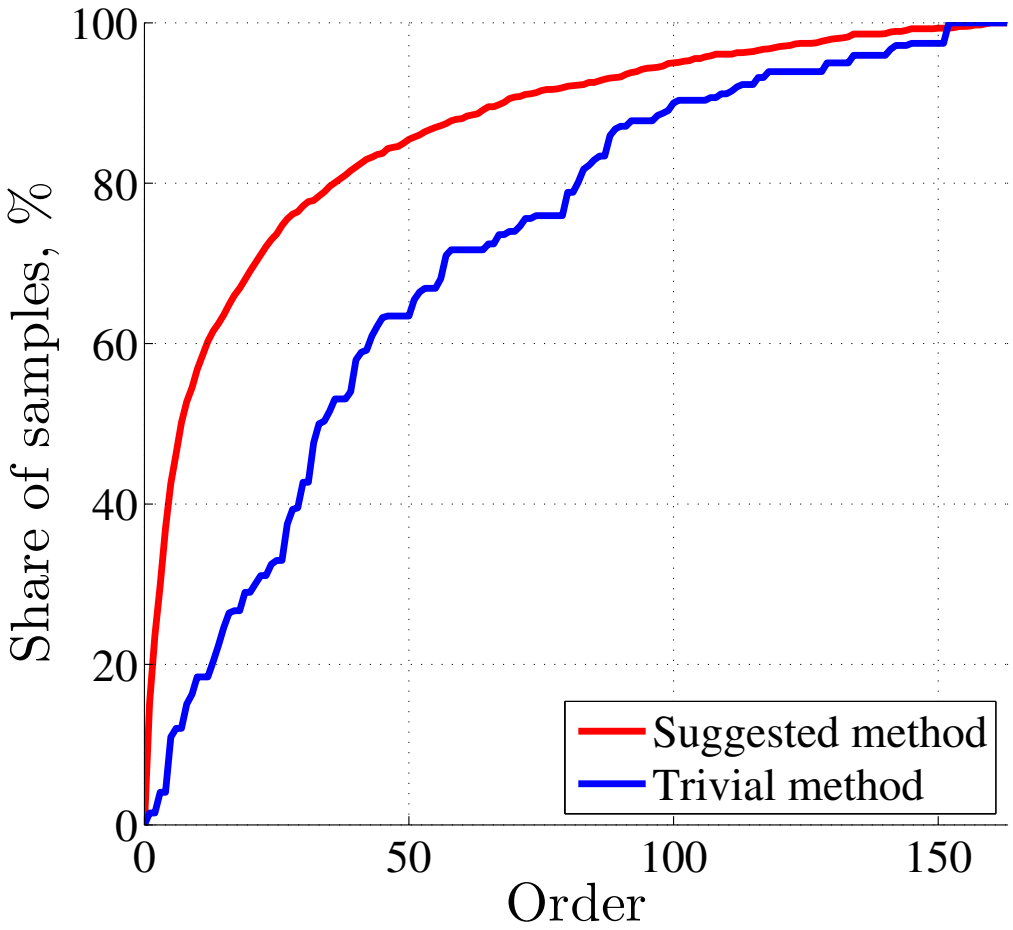
$$Q(R) = \frac{1}{|D|} \sum_{j=1}^{|D|} \mathrm{pos}\big(R(\mathbf{x}_j),\, z_{j,h}\big).$$

Критерий качества AUCH(R)

$\mathrm{AUC\ CH}(R) \in [0,\, 1]$ – площадь под кривой гистограммы $\#\{\mathrm{pos}(R(\mathbf{x}_j),\, z_{j,h}) \leq i\}$, где $z_{j,h}$ – номер экспертного кластера документа $\mathbf{x}_j$, а $i \in [1,\, k_h]$:

$$\mathrm{AUCH}(R) = \frac{1}{k_h|D|} \sum_{i=1}^{k_L} \#\{\mathrm{pos}\big(R(\mathbf{x}_j),\, z_{j,h}\big) \leq i\}.$$

Пример огибающей кумулятивной гистограммы



Предобработка документов

$W = \{w_1, \ldots, w_n\}$ – словарь.
Для построения W производится:

- Удаление стоп-слов
- Нормализация слов в документах
- Удаление часто (редко) встречающихся слов

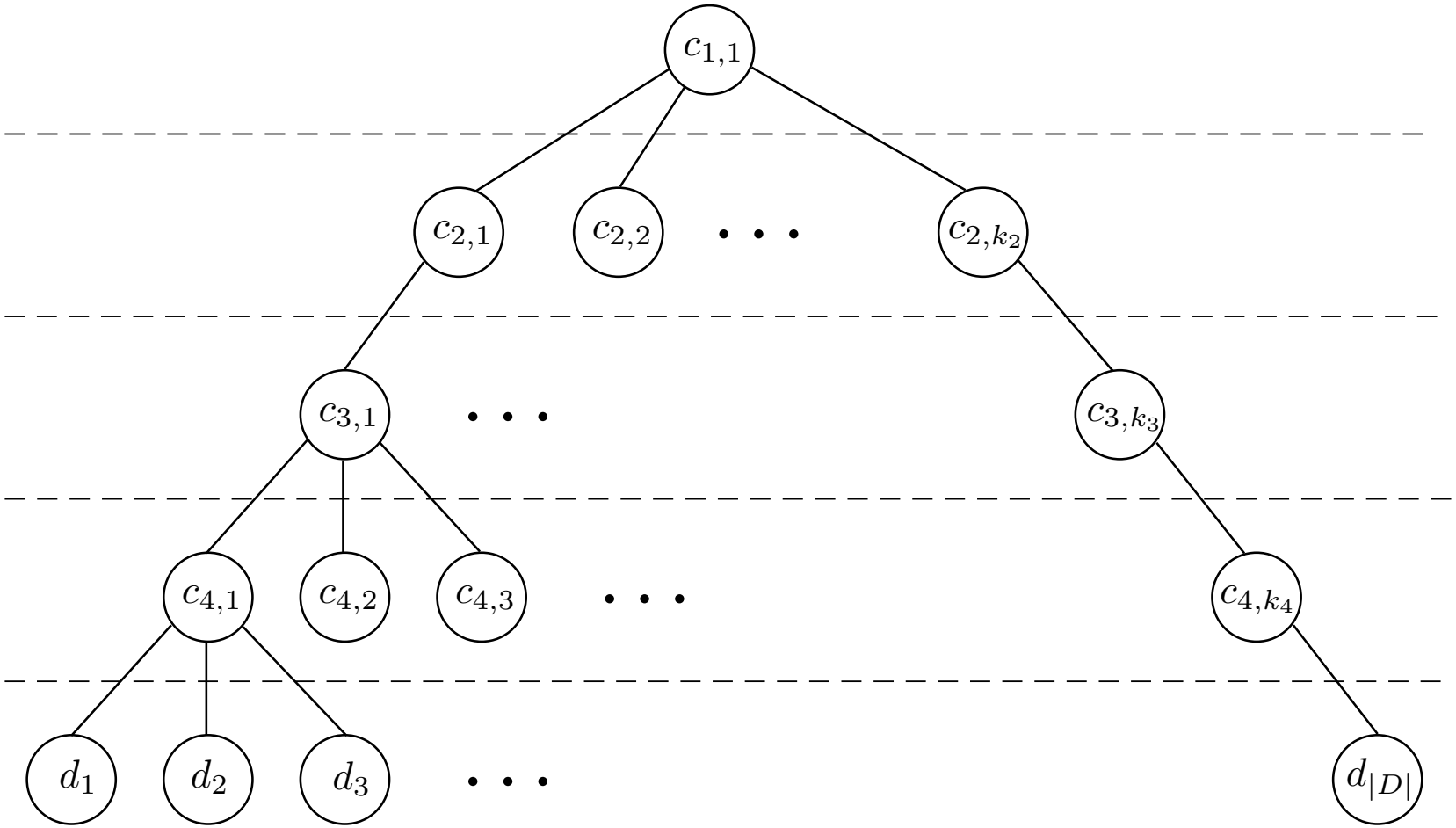
Документ — мешок слов

Документ d из коллекции D – неупорядоченный набор слов из словаря W, $d = \{w_j\}$, $j \in \{1, \ldots, n\}$.

Пусть $\mathbf{x}(d) \in \mathbb{R}^{|W|}$ – векторное представление документа d.

Иерархическая тематическая модель

Иерархическая структура $\mathcal{C}$ тематической модели M.



- $c_{l,i}$ –кластер уровня l с порядковым номером i на уровне l.
- $h = 4$ – нижний уровень кластеров.

Построение метрики

Способы представления документа

1) булево значение $[N(w_i, d) > 0]$,
2) значение произведения $\text{tf}(w_i, d) \cdot \text{idf}(w_i, D)$,
3) число слов w_i в документе $N(w_i, d)$,
4) doc2vec, сверточные нейронные сети.

Функция расстояния. Взвешенная метрика Минковского с фиксированным параметром $p \geq 1$:

$$\rho(\boldsymbol{\theta}, \mathbf{x}, \mathbf{y}) = \sqrt[p]{\sum_{i=1}^{n} \theta_i |x_i - y_i|^p}, \quad \text{где} \quad \sum_{i=1}^{n} \theta_i = 1, \quad \theta_1, \ldots, \theta_n \geq 0.$$

Функция сходства. Взвешенная косинусная функция сходства:

$$s(\mathbf{x}, \mathbf{y}) = \frac{\mathbf{x}^\mathsf{T} \boldsymbol{\Lambda} \mathbf{y}}{\sqrt{\mathbf{x}^\mathsf{T} \boldsymbol{\Lambda} \mathbf{x}} \sqrt{\mathbf{y}^\mathsf{T} \boldsymbol{\Lambda} \mathbf{y}}}, \quad \text{где} \quad \boldsymbol{\Lambda} = \text{diag}(\lambda_1, \cdots, \lambda_n).$$

Функция качества метрики

Пусть $\mathcal{P}(\mathbf{x})$ – множество документов из класса $c(\mathbf{x})$,
а $\overline{\mathcal{P}}(\mathbf{x})$ – множество документов остальных классов. Расстояние до k ближайших соседей из множества $\mathcal{P}$:

$$r_k(\mathbf{x}, \mathcal{P}) = \min_{\mathcal{A} \subset \mathcal{P}: |\mathcal{A}| = k} \sum_{\mathbf{y} \in D_{\mathcal{A}}} \rho(\boldsymbol{\theta}, \mathbf{x}, \mathbf{y}),$$

Функция близости $s_\rho(\mathbf{x})$ документа $\mathbf{x}$ к документам своего класса задается как

$$s_\rho(\mathbf{x}) = \frac{\overline{r}_k(\mathbf{x}, \overline{\mathcal{P}}(\mathbf{x})) - r_k(\mathbf{x}, \mathcal{P}(\mathbf{x}))}{\overline{r}_k(\mathbf{x}, \overline{\mathcal{P}}(\mathbf{x})) + r_k(\mathbf{x}, \mathcal{P}(\mathbf{x}))}.$$

Качество метрики ρ задается как

$$V(\rho, \boldsymbol{\theta}, D) = \frac{1}{|D|} \sum_{\mathbf{x} \in D} s_\rho(\mathbf{x}) \to \max_{\boldsymbol{\theta}}.$$

Построение оптимальной метрики

Пусть $\mathcal{J} = \{1, \ldots, |W|\}$ множество индексов всех признаков.
$\mathcal{A}$ – множество активных признаков. Инициализация:

$$\boldsymbol{\theta} = \mathbf{0}, \quad \mathcal{A} = \varnothing$$

1. Найти признак $\hat{i} \in \mathcal{J} \setminus \mathcal{A}$, доставляющий максимум функции качества $V(\rho, \hat{\boldsymbol{\theta}}, D_{\mathcal{L}})$:

$$\hat{i} = \arg\max_{i \in \mathcal{J} \setminus \mathcal{A}} V\left(\rho, \hat{\boldsymbol{\theta}}, D_{\mathcal{L}}\right),$$

$$\widehat{\boldsymbol{\theta}} = \arg\max_{i \in \left(\mathcal{J} \setminus (\mathcal{A} \cup \{\hat{i}\})\right)} V\left(\rho, \hat{\boldsymbol{\theta}}, D_{\mathcal{T}}\right), \quad \text{при условии} \quad \|\hat{\boldsymbol{\theta}}\| = 1,$$

2. Добавить найденный индекс $\hat{i}$ в множество $\mathcal{A}$. Если качество V изменилось значимо, вернуться на шаг 1.

Построение тематической модели с помощью найденной метрики

M – экспертная модель, $\hat{M}$ – построенная модель.
Ошибка классификации документа $\mathbf{x}$ задается как

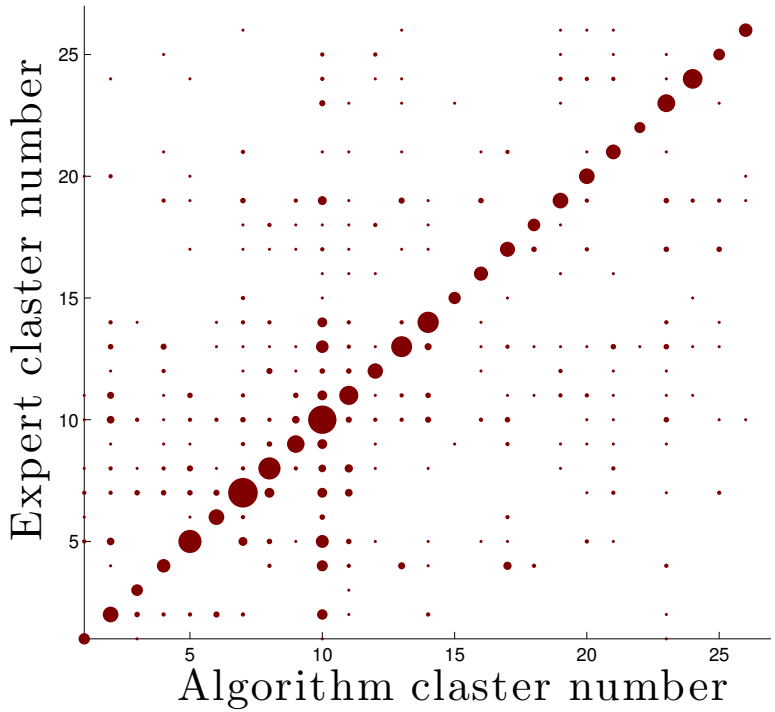
$$\upsilon(\mathbf{x}, M, \hat{M}) = \sum_{l=1}^{h} \rho\left(\boldsymbol{\mu}\left(B^{h-l}(c(\mathbf{x}))\right), \boldsymbol{\mu}\left(B^{h-l}(\hat{c}(\mathbf{x}))\right)\right).$$

Расстояние $\Upsilon(M, \hat{M})$ определяется как

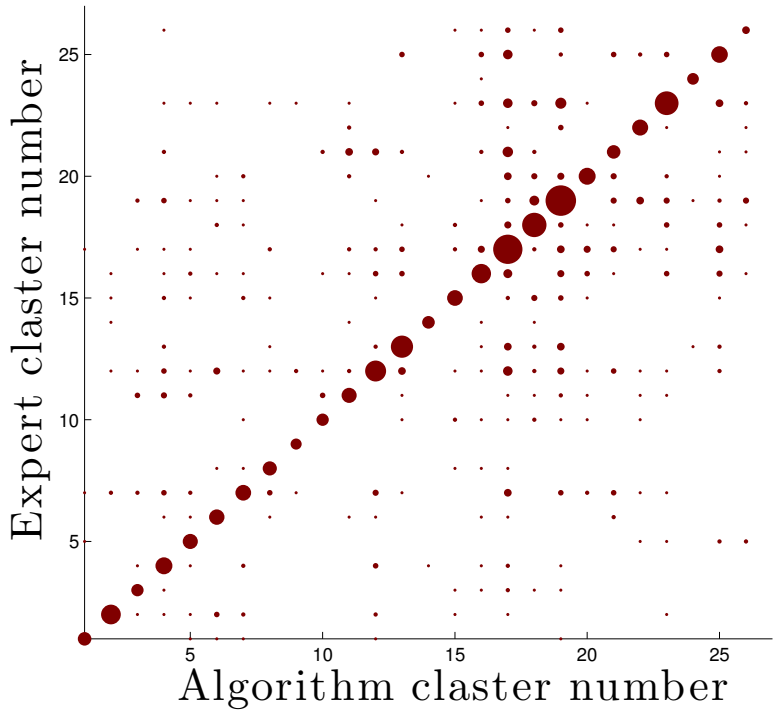
$$\Upsilon(M, \hat{M}) = \sum_{\mathbf{x} \in D} \sum_{l=1}^{h} \upsilon(\mathbf{x}, M, \hat{M}).$$

	Булевы признаки	$\mathrm{tf} \cdot \mathrm{idf}$	Целое число
Υ	**398**	710	771
$\Upsilon_{\mathcal{A}}$	**364**	630	650

Сравнение



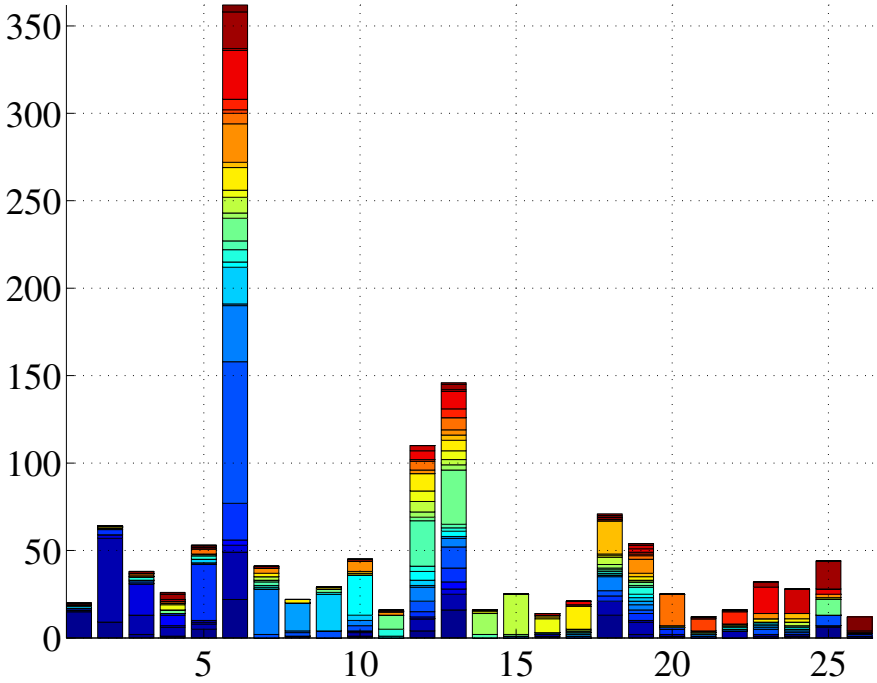
Булевы признаки,
без оптимизации метрики.



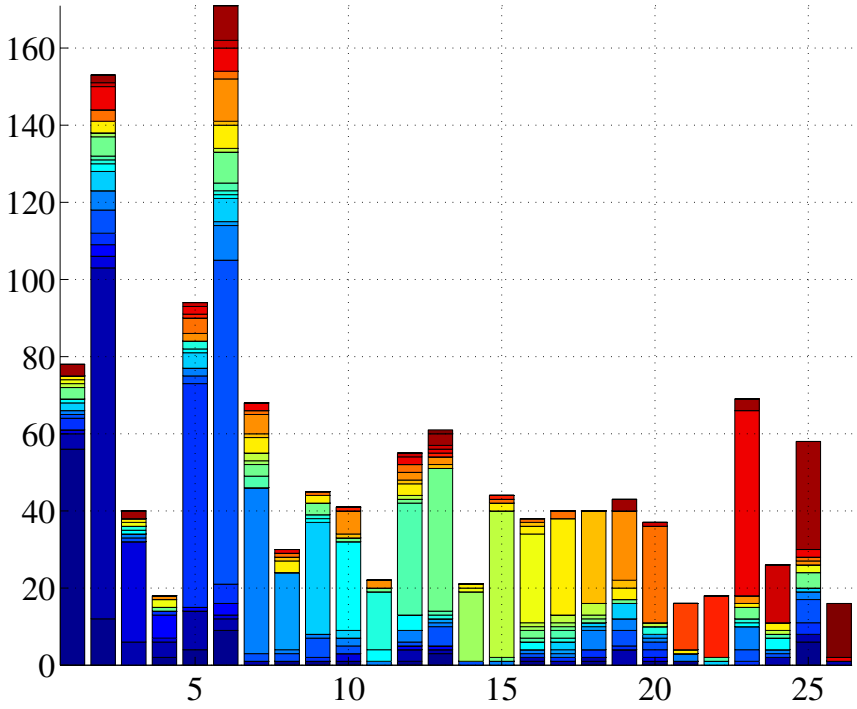
Булевы признаки,
с оптимизацией метрики.

Точка с координатами (x, y) – число документов с экспертным кластером x и алгоритмическим кластером y.

Сравнение



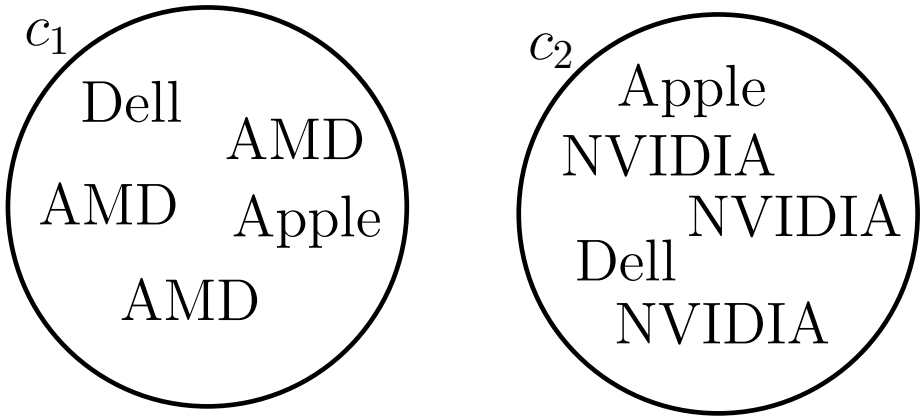
$\mathrm{tf} \cdot \mathrm{idf}$,
с оптимизацией метрики.



Булевы признаки,
с оптимизацией метрики.

Точка с координатами (x, y) – число документов с экспертным кластером x и алгоритмическим кластером y.

Важность слова при кластеризации



$W = \{\text{Dell, Apple, AMD, NVIDIA}\}$

$$\boldsymbol{\mu}(c_1) = [0.2, 0.2, 0.6, 0]^\mathsf{T},$$

$$\boldsymbol{\mu}(c_2) = [0.2, 0.2, 0, 0.6]^\mathsf{T}.$$

$\mathbf{p}_j$ – нормированный вектор j-ых компонент центров кластеров $\boldsymbol{\mu}(c_k)$:

$$\mathbf{p}_j = [\mu(c_1)_j, \ldots, \mu(c_K)_j]^\mathsf{T}, \quad \mathbf{p}_j \mapsto \frac{\mathbf{p}_j}{\|\mathbf{p}_j\|}.$$

Энтропия $I(w_j)$ слова w_j : $I(w_j) = \sum_{k=1}^{K} -p_{jk} \log(p_{jk})$.

В случае нашего примера

$$\mathbf{P} = \begin{pmatrix} 0.5 & 0.5 & 1.0 & 0 \\ 0.5 & 0.5 & 0 & 1.0 \end{pmatrix}^\mathsf{T} \qquad I = [0.69, 0.69, 0, 0]^\mathsf{T}$$

Модель оценки важности слов

Матрица важности слов

$$\mathbf{\Lambda} = \mathrm{diag}(\lambda_1, \ldots, \lambda_{|W|}).$$

Важность λ_j слова w_j:

Плоский случай

$$\lambda_j = 1 + \alpha \log\bigl(1 + I(w_j)\bigr).$$

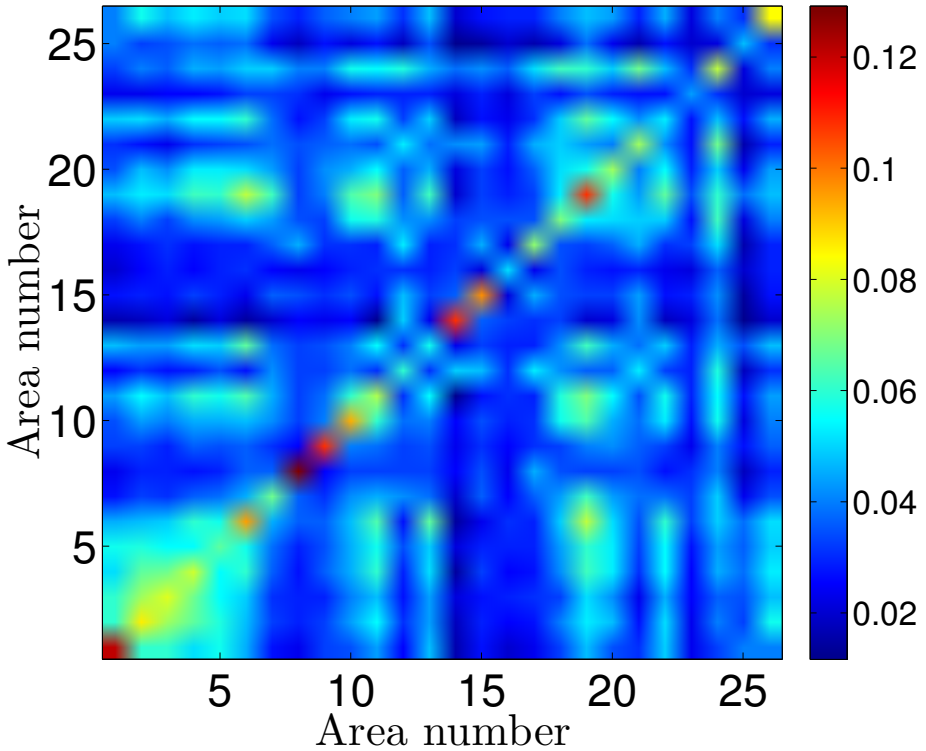
$$\alpha^* = \arg\min_{\alpha} \mathrm{AUCH}(R).$$

Иерархический случай

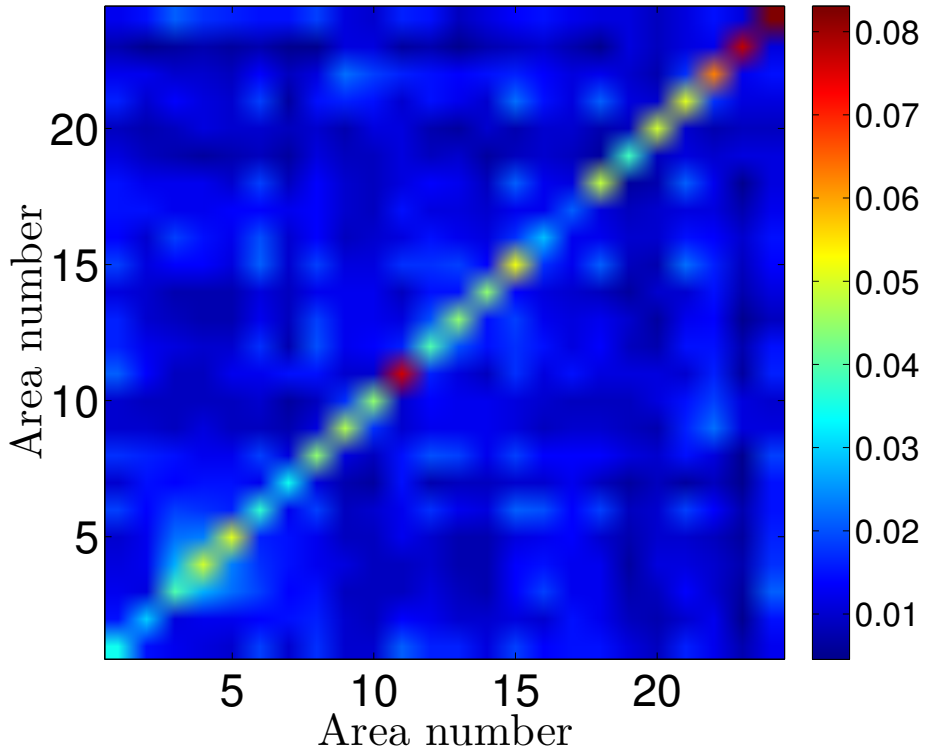
$$\lambda_j = 1 + \sum_{l=1}^{h} \alpha_l \log\bigl(1 + I^l(w_j)\bigr).$$

$$[\alpha_1^*, \ldots, \alpha_h^*] = \arg\min_{\alpha_1, \ldots, \alpha_h} \mathrm{AUCH}(R).$$

Сходство экспертных кластеров



$\mathbf{\Lambda} = \mathbf{I}$



$\mathbf{\Lambda} = \mathbf{\Lambda}^*$

Цвет точки (x, y) на графике соответствует значению сходства между кластерами с номерами x и y.

Иерархическая взвешенная функция сходства

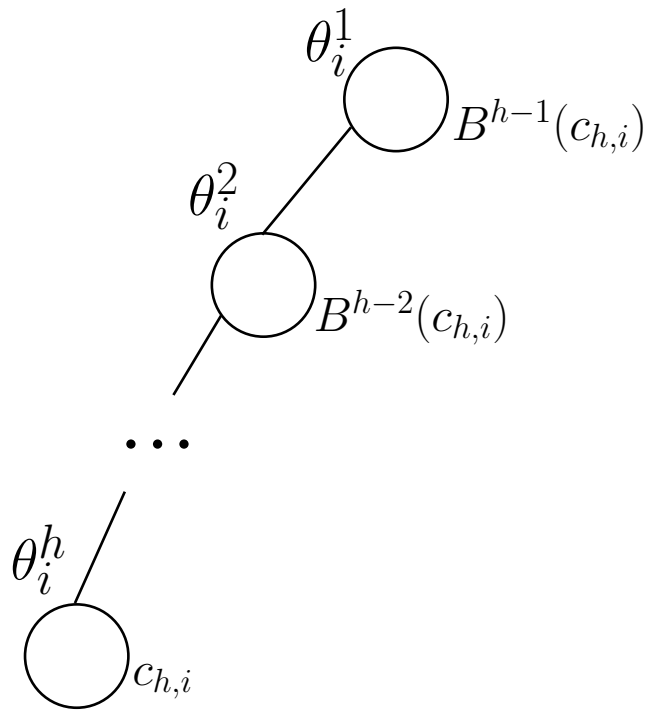
Сходство $s(\cdot,\cdot)$ документа $\mathbf{x}$ и кластера $c_{\ell,i}$ на уровне ℓ называется:

$$s(\mathbf{x},\ c_{\ell,i}) = \mathbf{x}^{\mathsf{T}}\mathbf{\Lambda}\boldsymbol{\mu}(c_{\ell,i}).$$

Иерархическое взвешенное сходство документа $\mathbf{x}$ и кластера $c_{\ell,h}$ называется

$$s_h(\mathbf{x},\ c_{h,i}) = \sum_{\ell=1}^{h} \theta_i^{\ell} s\big(\mathbf{x},\ B^{h-\ell}(c_{h,i})\big),$$

где θ_i^{ℓ} значимость уровня ℓ, а $B^{h-\ell}$ – оператор, возвращающий для кластера $c_{h,i}$ его родительский кластер уровня ℓ.

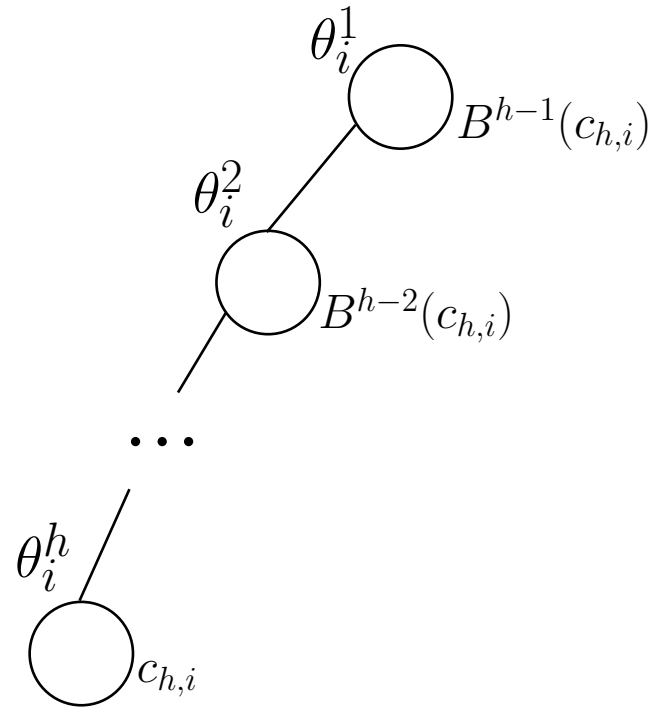


Оптимизация параметров

Поиск оптимальных $\boldsymbol{\theta}_i$ сводится к решению следующей задачи:

$$\boldsymbol{\theta}_i^* = \arg\max_{\boldsymbol{\theta}_i} \sum_{\mathbf{x} \in c_{h,i}} \sum_{\ell=1}^{h} \theta_i^\ell s\big(\mathbf{x},\, B^{h-\ell}(c_{h,i})\big) + \psi \sum_{\ell=1}^{h} \left(\theta_i^\ell - \frac{1}{h}\right)^2,$$

$$\sum_{\ell=1}^{h} \theta_i^\ell = 1, \quad \theta_i^\ell \geq 0, \quad \ell \in \{1 \ldots h\}.$$



Итеративный алгоритм

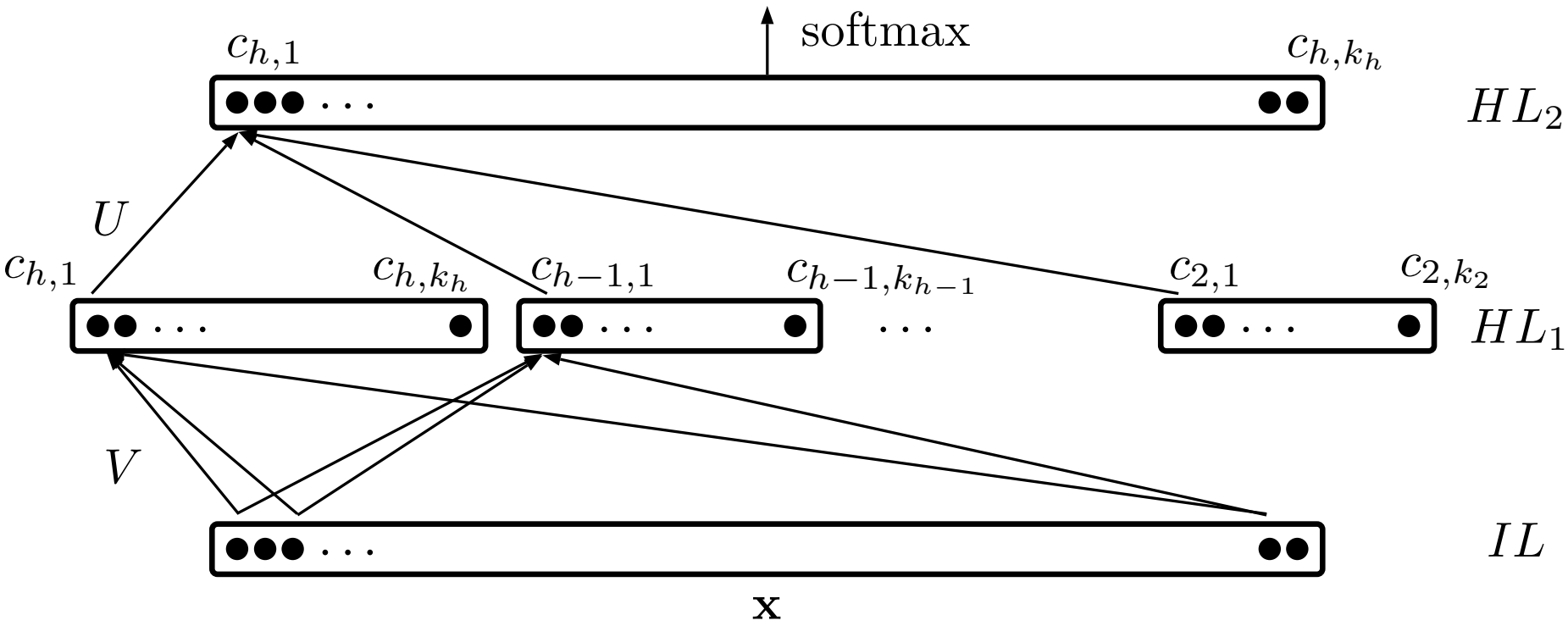
1) Найти $\boldsymbol{\mu}(c_{l,i})$ и оптимальную $\boldsymbol{\Lambda}$ при фиксированных $\boldsymbol{\theta}_i$;
2) Найти оптимальные $\boldsymbol{\theta}_i$ при фиксированной $\boldsymbol{\Lambda}$;

Критерий остановки: $\Delta\boldsymbol{\theta}_i$ и $\Delta\boldsymbol{\Lambda}$ малы на нескольких последовательных итерациях.

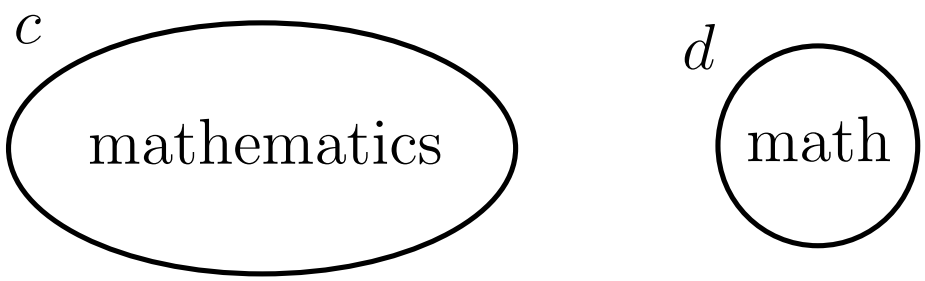
Представление иерархической функции сходства в виде нейронной сети.

Каждому кластеру $c_{l,i}$ в матрице $\mathbf{V}$ соответствует строка $\mathbf{v}_j^{\mathsf{T}}$ с номером j

$$\mathbf{v}_j = \mathbf{\Lambda}\boldsymbol{\mu}(c_{l,i}), \qquad j = i + \sum_{p=l-1}^{h} k_p,$$



Векторное представление слов



$$W = \{\text{mathematics, math}\}$$
$$\boldsymbol{\mu}(c) = [1,\ 0]^{\mathsf{T}},$$
$$\mathbf{x}(d) = [0,\ 1]^{\mathsf{T}}.$$

word2vec

$\mathbf{w}(\text{“mathematics”})^{\mathsf{T}}\mathbf{w}(\text{“math”}) = 0.7.$

функция сходства

$\boldsymbol{\mu}(c)^{\mathsf{T}}\boldsymbol{\Lambda}\mathbf{x}(d) = 0.$

Идея: для каждого слова из кластера искать наиболее близкое слово из документа.

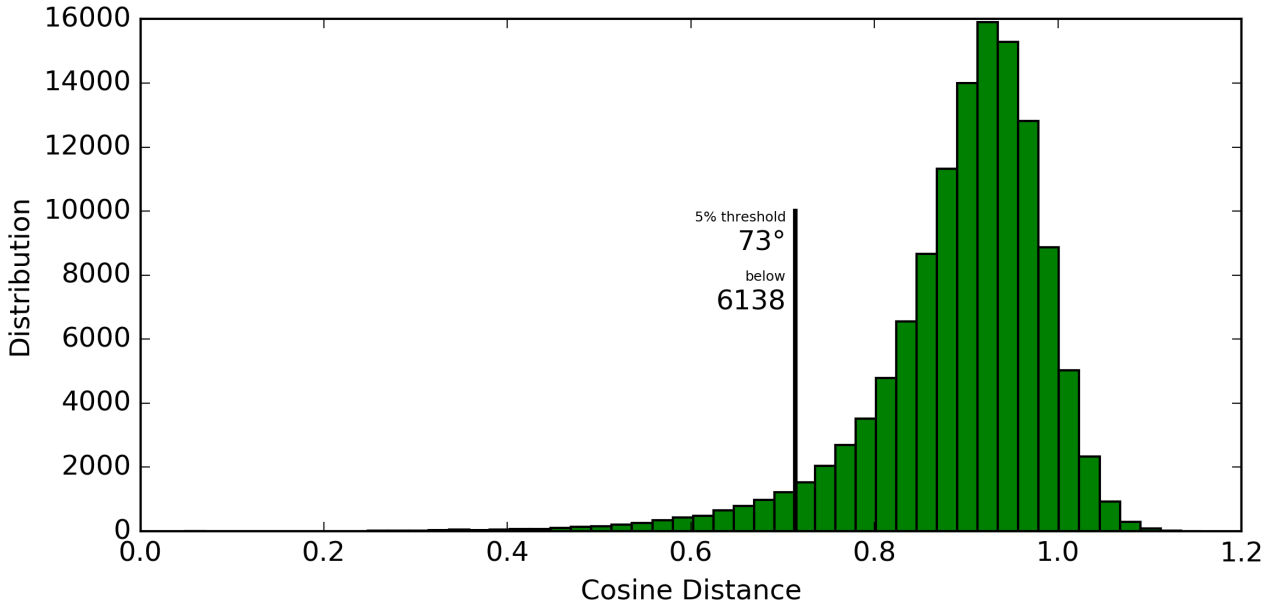
Для каждого слова $w_{m_1} \in W$ ищется наиболее близкое к нему слово w_{m_2}. На позицию m_1 вектора $\boldsymbol{\mu}$ ставится значение скалярного произведения векторных представлений слов w_{m_1} и w_{m_2}:

$$\mu_m = \mathbf{w}(w_{m_1})^{\mathsf{T}}\mathbf{w}(w_{m_2}). \quad \boldsymbol{\mu}(c) \to [1,\ 0.7]^{\mathsf{T}}.$$

Устранение шумового сходства

Распределение $1 - cos(\mathbf{w}_1, \mathbf{w}_2)$ для 500 тем

источник: http://www.trivial.io/word2vec-on-databricks/



$$\mu'_m \mapsto \begin{cases} f(\mu_m) = \left(1 + \cos\left(-\pi\left(1.5 - \mu_m\right)\right)\right)^p, & \text{если } \mu_m \geq 0.5 \\ 0, & \text{иначе.} \end{cases}$$

Учет важности слов:

$$\mu''_m \mapsto \mu'_m \cdot \lambda_m.$$

Существующие методы, Supervised PLSA (ARTM)

Строится регуляризированная вероятностная тематическая модель SuHiPLSA:

$$\boldsymbol{\Phi}^*, \boldsymbol{\Theta}^* = \arg\max_{\boldsymbol{\Phi},\boldsymbol{\Theta}} L(\boldsymbol{\Phi}, \boldsymbol{\Theta}) + \lambda\Omega(\boldsymbol{\Theta}, \mathbf{Z}),$$

$$\Omega(\boldsymbol{\Theta}, \mathbf{Z}) = -\|\boldsymbol{\Theta} - \mathbf{Z}\|_1.$$

где $\Omega(\boldsymbol{\Theta}, \mathbf{Z})$ – штраф за несоответствие прогнозных тем $\boldsymbol{\Theta}$ экспертным $\mathbf{Z}$.

Для учета иерархической структуры используется регуляризатор:

$$\Omega_h(\boldsymbol{\Theta}, \mathbf{Z}_2, \ldots, \mathbf{Z}_h) = \sum_{l=2}^{h} \sum_{d \in D} \sum_{c \in \{c_{l,i}\}} |z_{cd,l} - \theta_{cd,l}|, \quad \text{где}$$

$$\theta_{cd,l} = \frac{1}{|C_l(d)|} \sum_{c \in C_l(d)} \theta_{cd}, \quad C_l(d) = \{c : B^{h-l}(c) = B^{h-l}\big(c(d)\big)\}.$$

$C_l(d)$ – множество тем уровня h, у которых родительская тема на уровне l совпадает с родительской темой уровня l экспертной темы $c(d)$ документа d уровня h.

Существующие методы, hNB

Слово в документе зависит только от класса документа и не зависит от контекста и его позиции в документе:

$$\mathrm{P}(d|c_{h,j}) = \mathrm{P}(|d|) \prod_{w \in d} \mathrm{P}(w|c_{h,j}).$$

Вероятности слов $\mathrm{P}(w_i|c_{h,j}) = \theta_{ij}$ и априорные вероятности классов $\mathrm{P}(c_{h,j}) = \theta_{0j}$ оцениваются как

$$\hat{\theta}_{ij} = \frac{1 + \sum_{d \in D} N(w_i, d)\mathrm{P}(c_{h,j}|d)}{|W| + \sum_{w \in W} \sum_{d' \in D} N(w, d')\mathrm{P}(c_{h,j}|d')}, \qquad \hat{\theta}_{0j} = \frac{1}{|D|} \sum_{d \in D} \mathrm{P}(c_{j,h}|d),$$

$$\mathrm{P}(c_{h,j}|d) = [c(d) = c_{h,j}].$$

Усреднение параметров с родительскими кластерами

$$\hat{\theta}_{ij} = \lambda_{j,1}\hat{\theta}_{ij,1} + \lambda_{j,2}\hat{\theta}_{ij,2} + \ldots + \lambda_{j,h}\hat{\theta}_{ij,h}, \qquad \sum_{i=1}^{h} \lambda_{j,i} = 1,$$

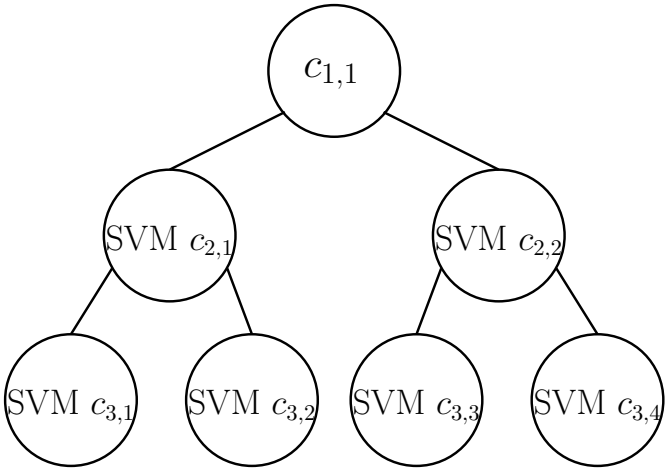
Существующие методы, hSVM

Для каждого кластера $c_{\ell,i}$ уровня ℓ обучается двухклассовый SVM по принципу “один против всех”.

Вероятность оценивается с помощью метода Платта (Platt, 2000):

$$P(c_{l,i}|\mathbf{x}) = \frac{1}{1+\exp(A\cdot\hat{m}(\mathbf{x})+B)},$$

где $\hat{m}(\mathbf{x})$ – отступ объекта $\mathbf{x}$,
A, B – числовые параметры.

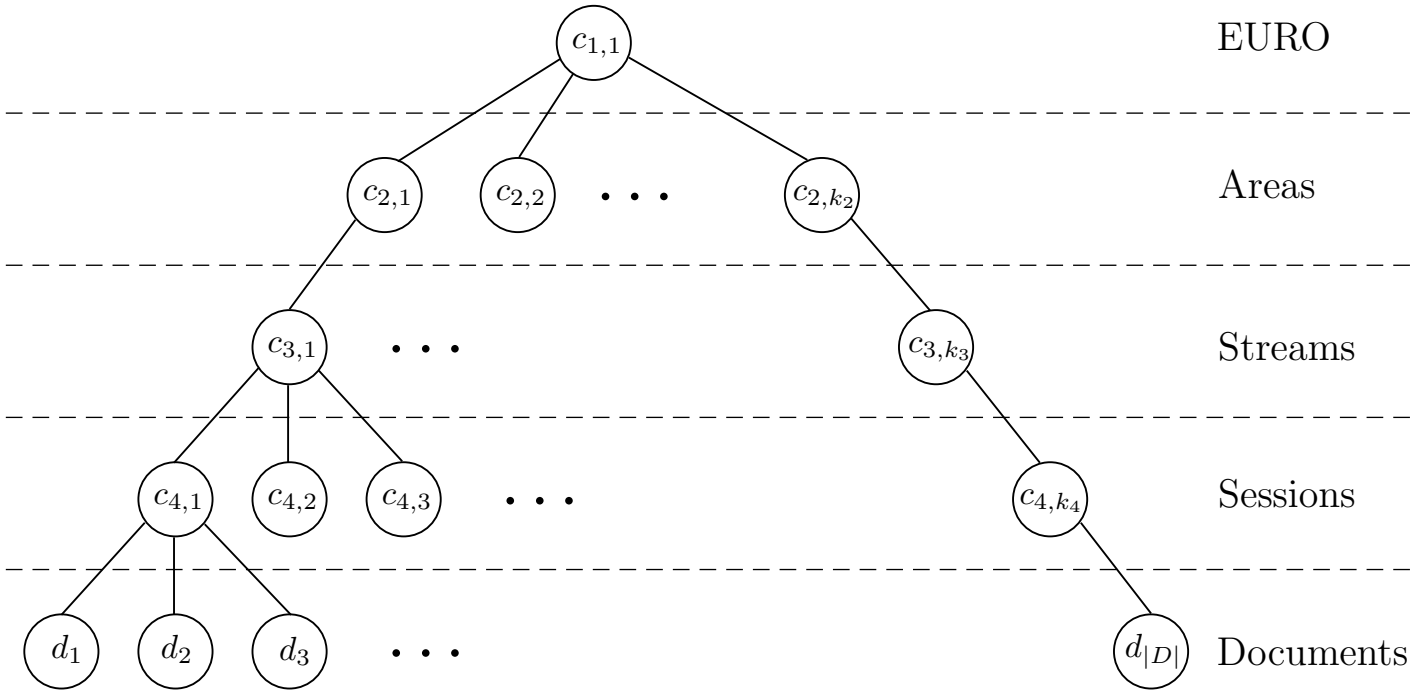


Пусть для некоторого объекта $\mathbf{x}$ вероятности $p(c_{2,1}) > p(c_{2,1})$, $p(c_{3,1}) > p(c_{3,2})$, $p(c_{3,4}) > p(c_{3,3})$.

Оператор релевантности $R_{SVM}(\mathbf{x})$ вернет следующую перестановку:

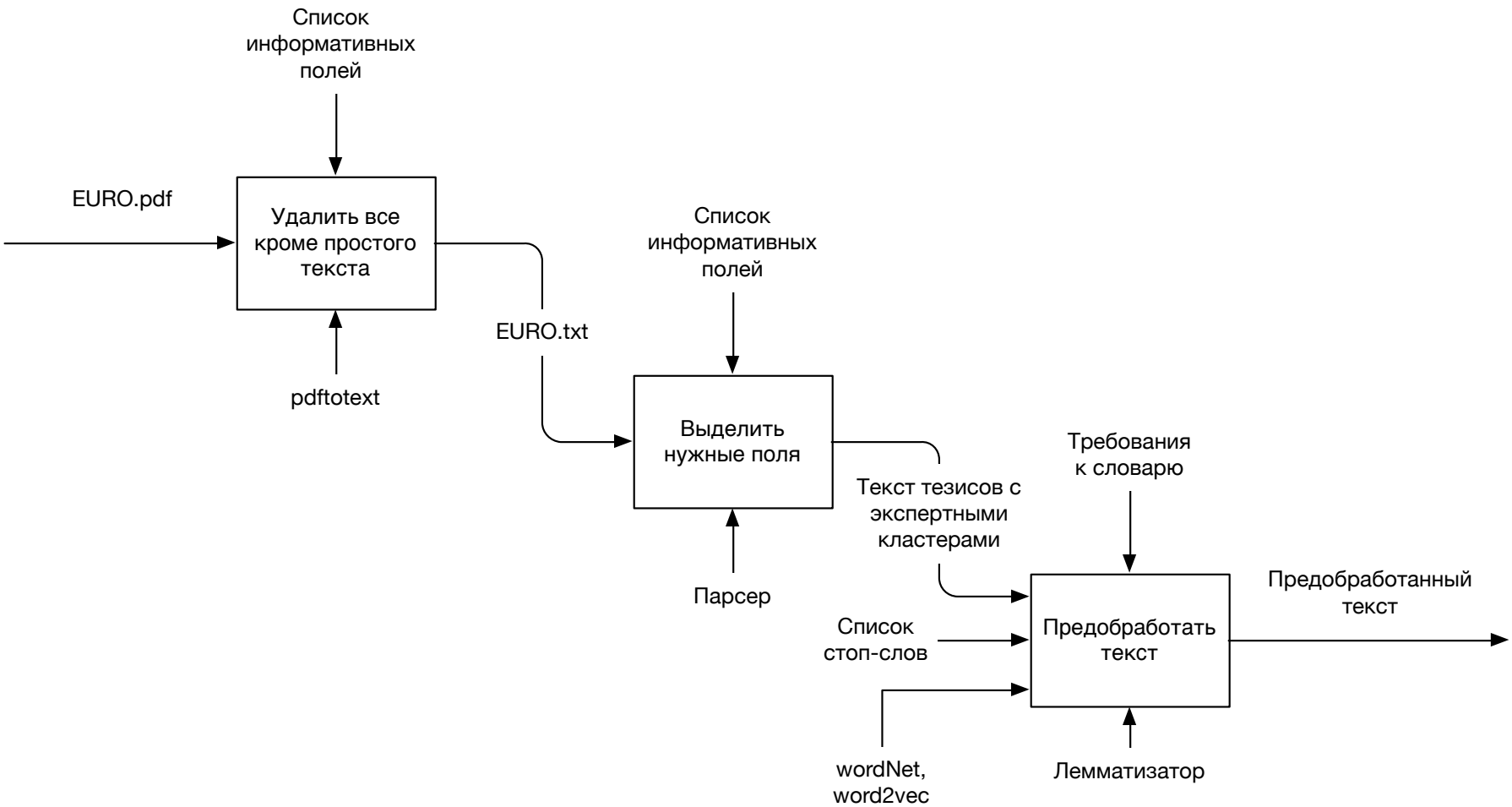
$$R_{SVM}(\mathbf{x}) = (\overbrace{1,2}^{c_{2,1}}, \overbrace{4,3}^{c_{2,2}})$$

Построение тематической модели конференции EURO

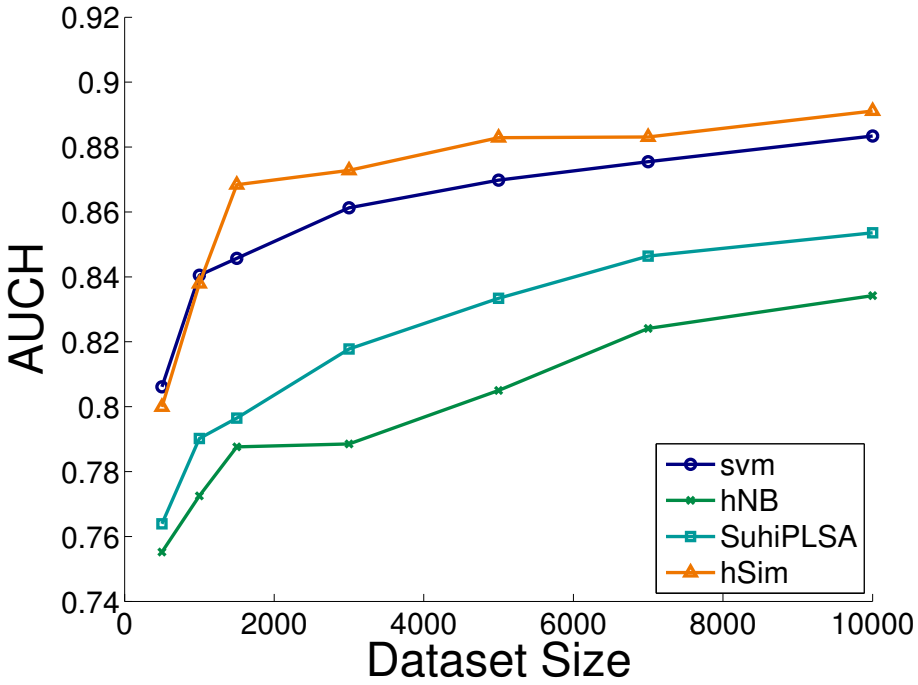


1) размер коллекции $|D| = 15527$ документов,
2) размер словаря $|W| = 24304$ слова,
3) число кластеров второго уровня (Area) $k_2 = 26$,
4) число кластеров третьего уровня (Stream) $k_3 = 264$.

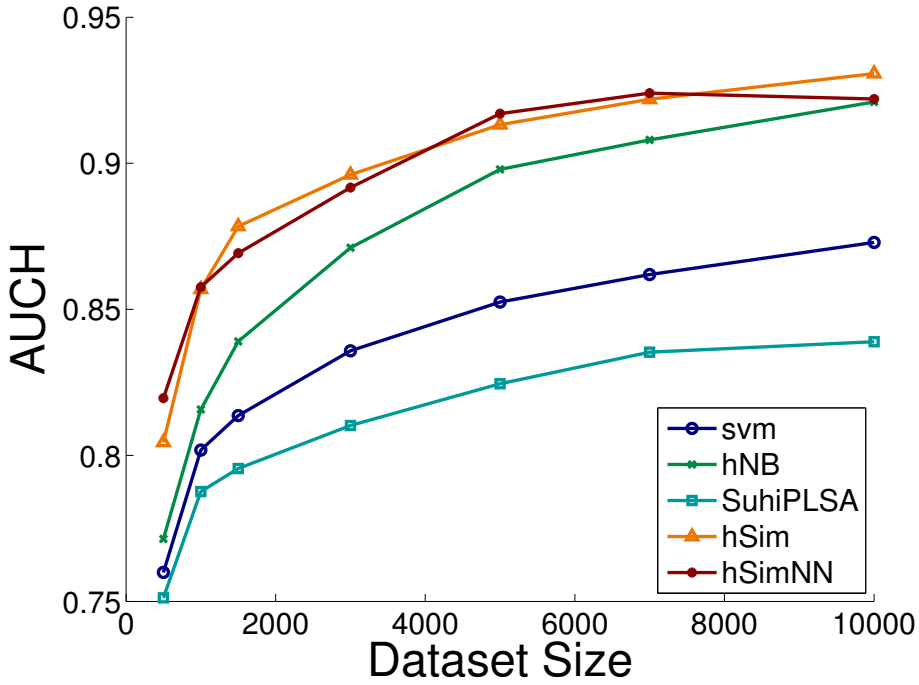
Процесс предобработки документов



Результаты классификации при различном числе обучающих документов



Плоская классификация



Иерархическая классификация

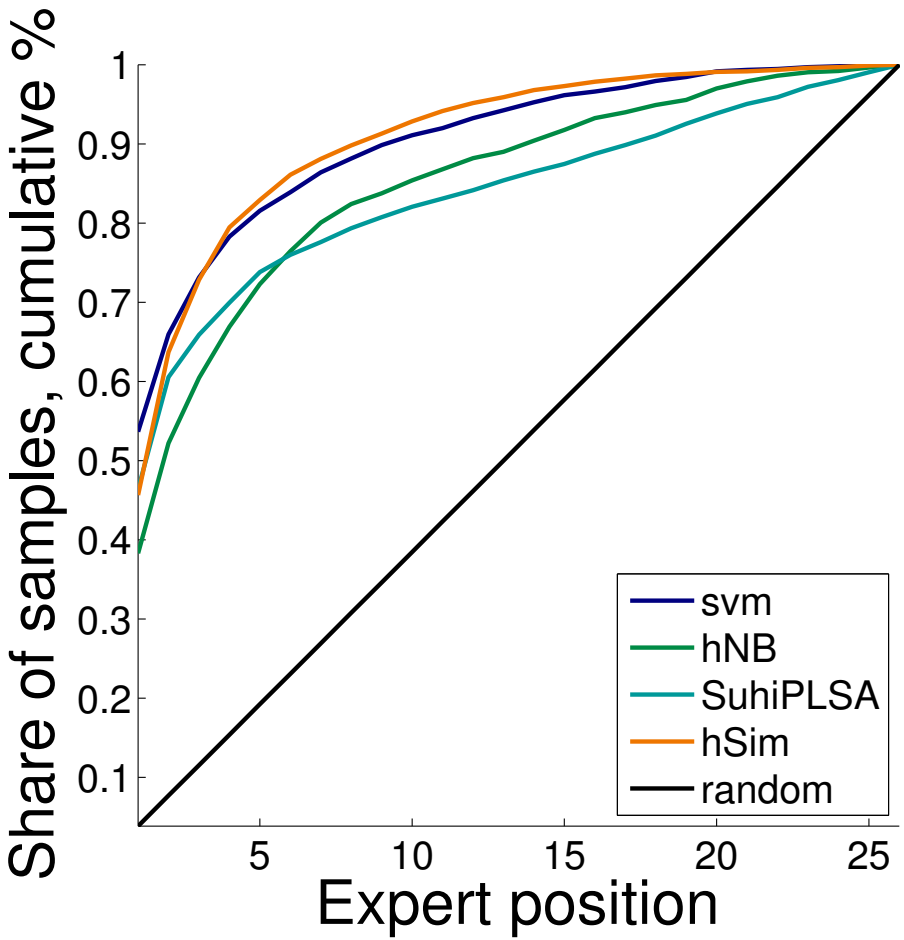
Зависимость значений AUCH операторов релевантности $R(\cdot)$, построенных с помощью алгоритмов svm, hNB, suhiPLSA, hSim и hSimNN, от размера обучающей выборки.

Результаты классификации при различном числе обучающих документов

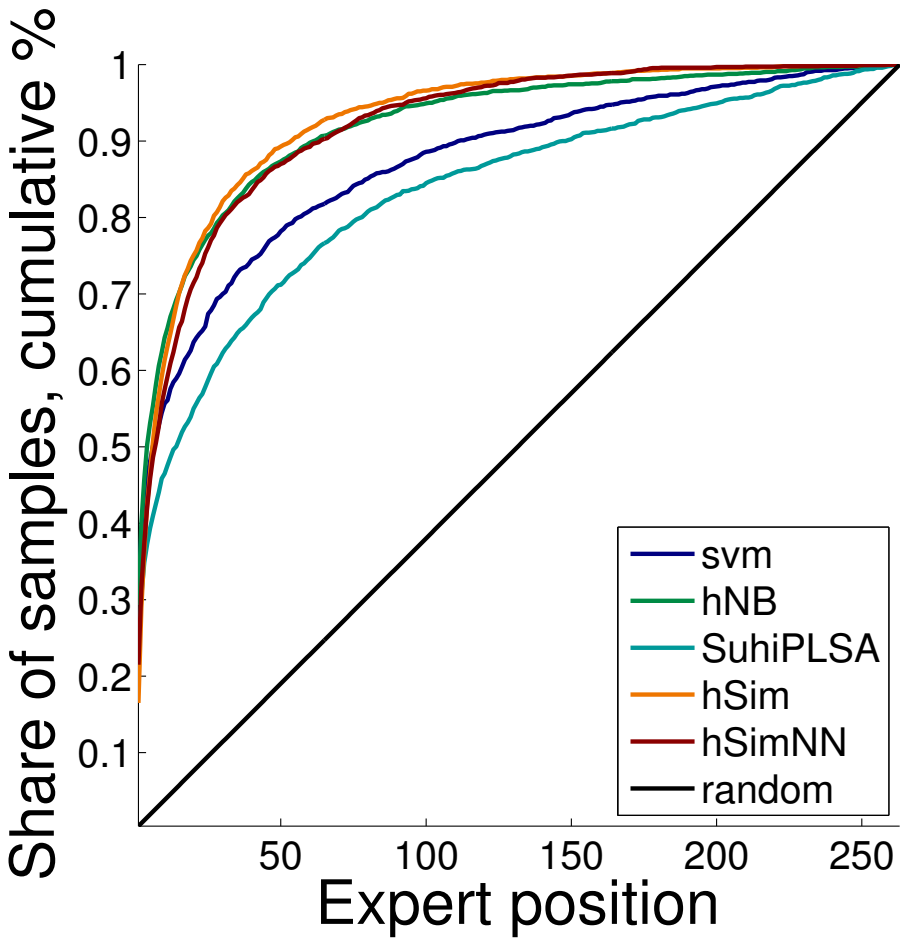
Размер выборки	500	1000	1500	3000	5000	7000	10000
svm	0.76	0.80	0.81	0.84	0.85	0.86	0.87
hNB	0.77	0.82	0.84	0.87	0.90	0.91	0.92
suhiPLSA	0.75	0.79	0.80	0.81	0.82	0.84	0.84
hSim	0.80	**0.86**	**0.88**	**0.90**	**0.91**	**0.92**	**0.93**
hSimNN	**0.82**	**0.86**	0.87	**0.89**	**0.92**	**0.92**	0.92

Значения функционалов качества AUCH на уровне Stream для операторов релевантности, построенных с помощью сравниваемых алгоритмов.

Сравнение AUCH для обучающей выборки из 10000 документов

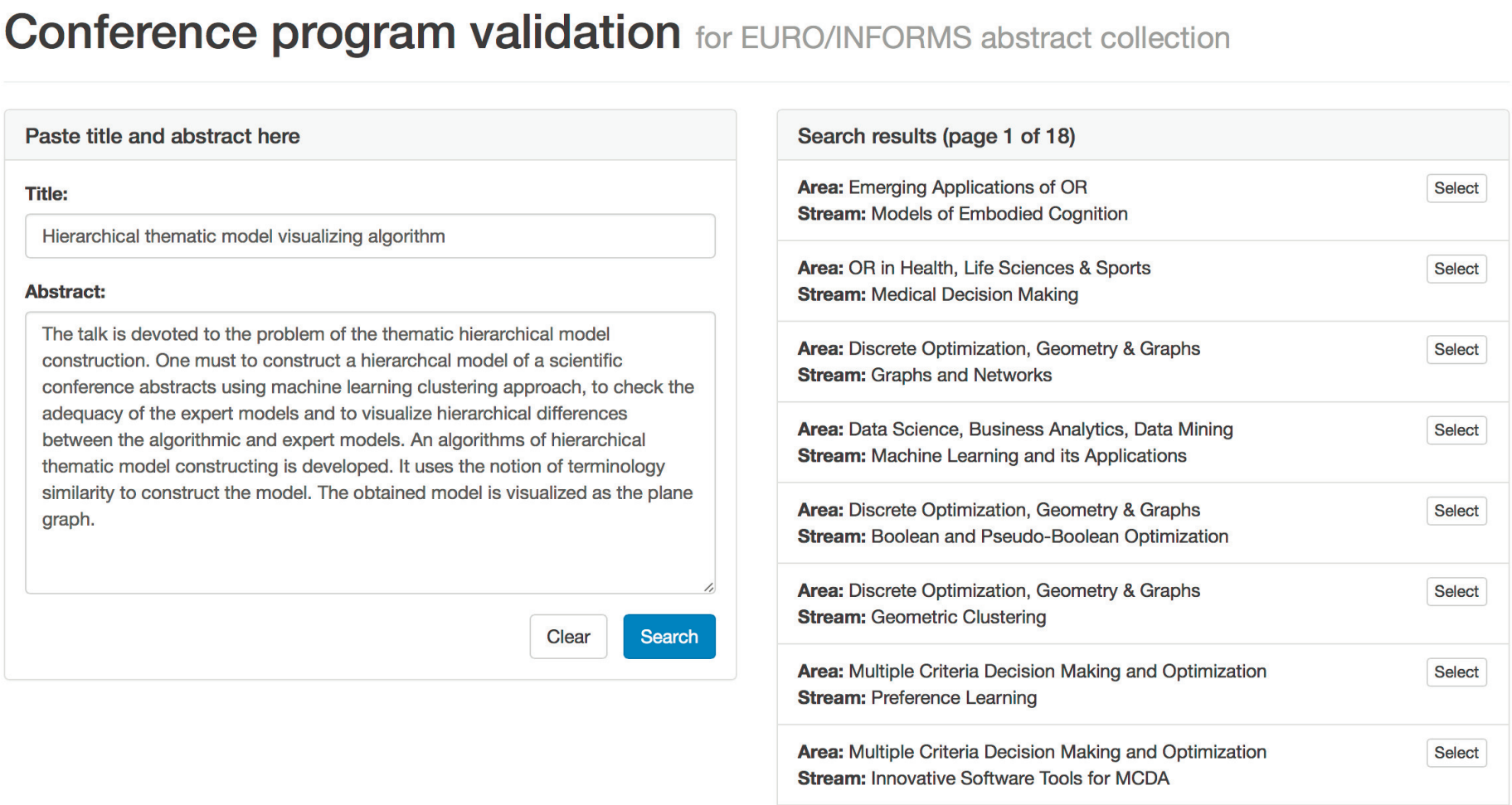


Плоская классификация.



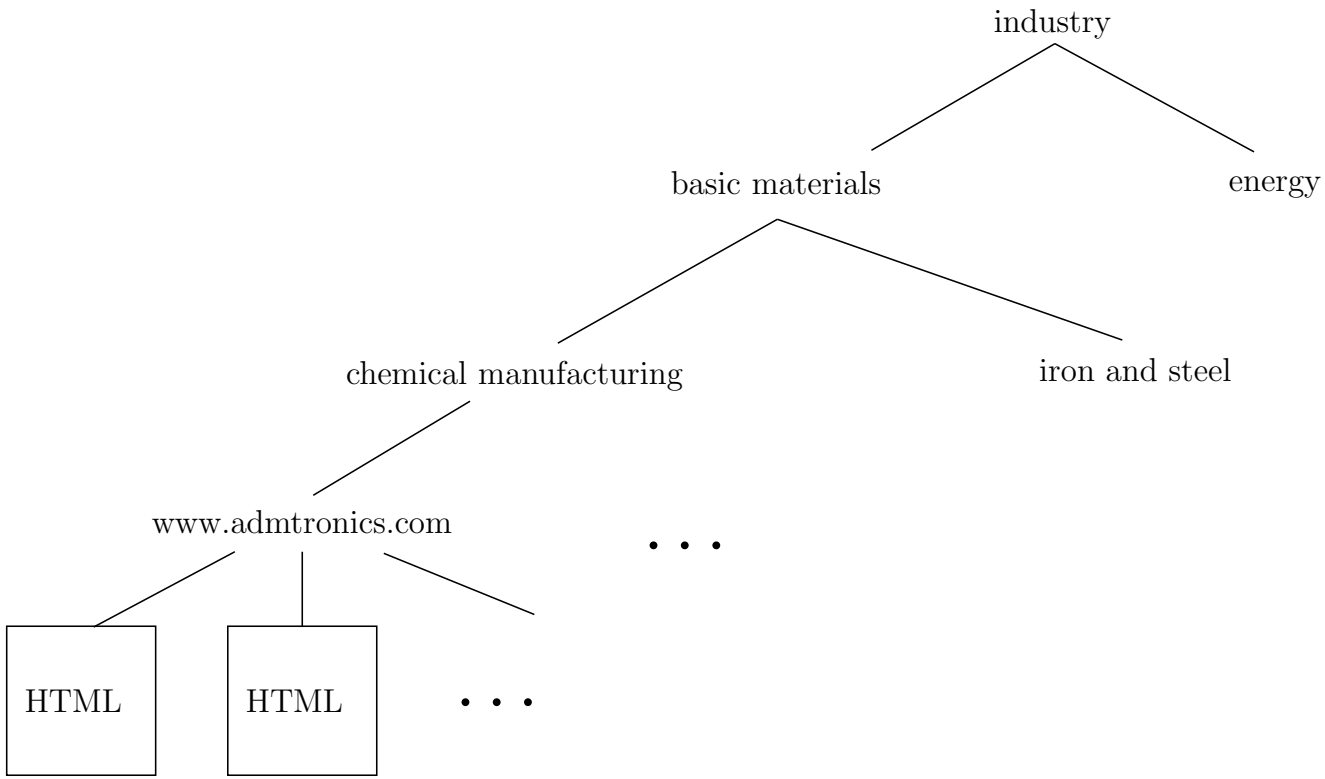
Иерархическая классификация.

Экспертная система

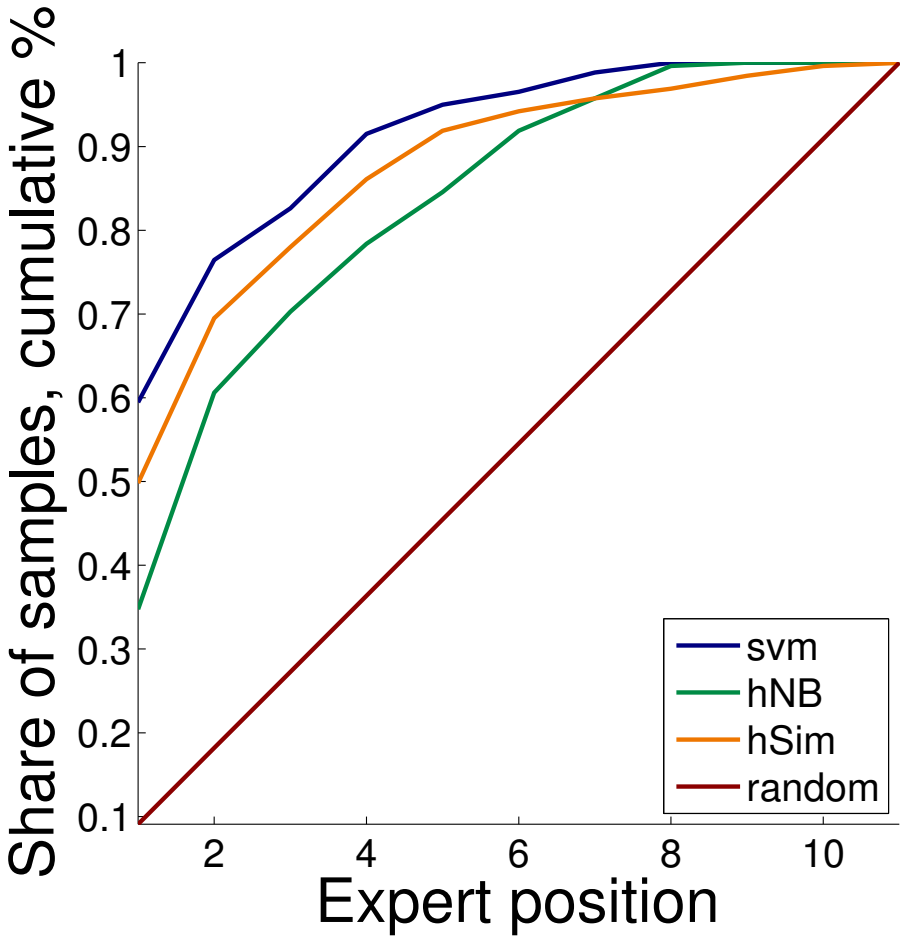


Экспертная система для поиска релевантных кластеров для неразмеченных документов.

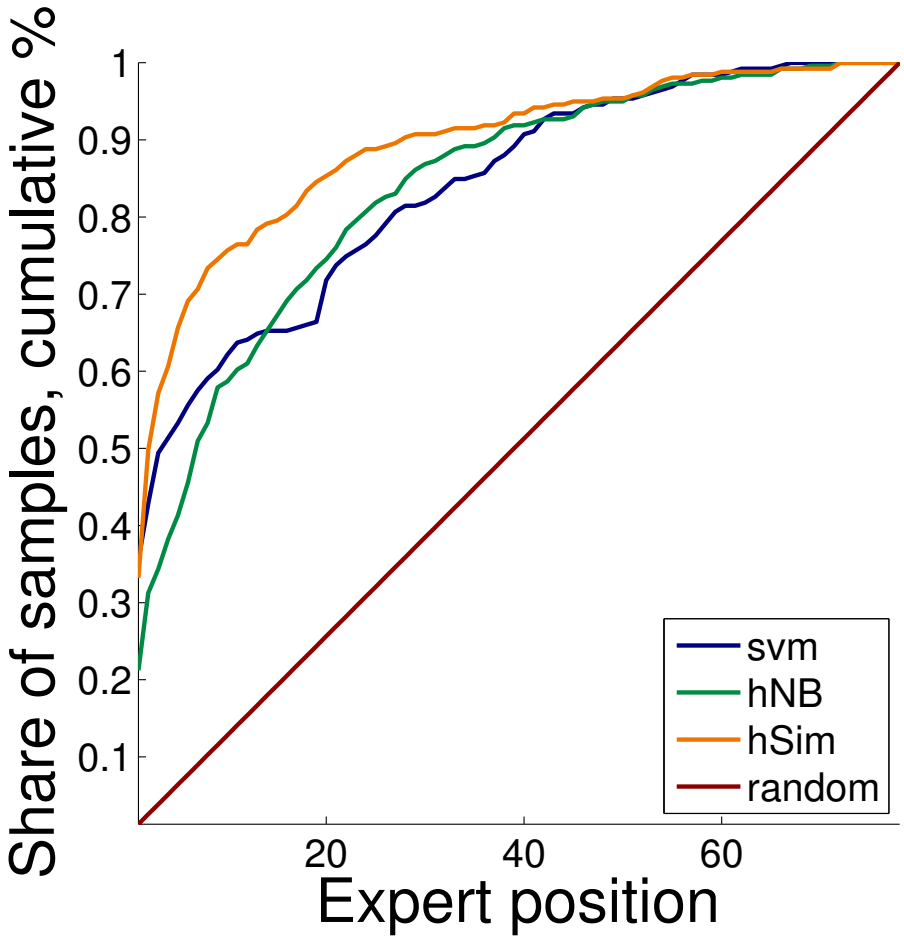
Построение тематической модели коллекции веб-сайтов индустриального сектора



1) размер коллекции $|D| = 1036$ документов,
2) размер словаря $|W| = 18775$ слова,
3) число кластеров второго уровня $k_2 = 11$,
4) число кластеров третьего уровня $k_3 = 78$.



Плоская классификация.



Иерархическая классификация.

Тип классификации	Плоская	Иерархическая
svm	**0.86**	0.83
hNB	0.80	0.83
hSim	0.83	**0.89**