Министерство образования и науки Российской Федерации
Московский физико-технический институт (государственный университет)
Факультет управления и прикладной математики
Кафедра «Интеллектуальные системы»
при Вычислительном центре им. А. А. Дородницына ФИЦ ИУ РАН

Митяшов Андрей Андреевич

Декомпозиция смеси распределений на основе эмпирических данных в задачах текстовой кластеризации

010656 – Математические и информационные технологии

Выпускная квалификационная работа магистра

Научный руководитель:
д.ф.-м.н.
Стрижов Вадим Викторович

Москва

2016 г.

Содержание

Аннотация

Данная работа посвящена задаче экспертно-интерпретируемой декомпозиции смеси распределений цен государственных закупок. Для оценки параметров распределений предлагается использовать текстовую кластеризацию с учетом дополнительных атрибутов. Кроме того, для каждой закупки предлагается оценить ее типичность и превышение относительно рыночной цены. Результаты данного исследования будут применены в проекте, посвященном анализу государственных закупок.

Ключевые слова: *декомпозиция смеси распределений, кластеризация коротких текстов, тематическое моделирование, DBScan, Hierarchical DBScan.*

1 Введение

Актуальность темы. Суммарный объем государственных закупок в Российской Федерации достигает 6 триллионов рублей за год. Однако, значительная часть этих средств (по некоторым данным вплоть до 15%) крадется, а не расходуется по назначению. Это происходит с помощью завышения цены закупки.
Однако, представляется возможным выявить нетипичные закупки и оценить процент превышения с помощью анализа базы данных государственных закупок. Для этого предлагается разбить всю базу на группы закупок, при этом каждая из групп интерпретируется как одна форма выпуска определенного товара. После этого в каждой группе можно выявить нетипичные закупки с превышением.
Данная задача представляет собой задачу кластеризацию, сводимую к экспертно-интерпретируемой декомпозиции смеси распределений.
Результаты работы предложенного исследования могут быть использованы для выявления нарушений организациями, контролирующими расход государственных средств.

Цель работы. Построить алгоритм разделения смеси распределений, результаты работы которого являются экспертно-интерпретируемыми. Алгоритм предлагается построить на основе многомодальной текстовой кластеризации.

Методы исследований. Для решения задачи текстовой кластеризации использовался алгоритм, основанный на DBScan. Для уменьшения размерности задачи предлагается предварительно подготовить набор выборок с помощью тематического моделирования. Для оценки качества работы алгоритма использовался принцип максимума правдоподобия с использованием экспертных ограничений.

Научная новизна. В данной работе предлагается двухуровневая модель кластеризации, учитывающая экспертные оценки.

Практическая ценность. Будет разработан модуль кластеризации коротких (не более 1024 символов) текстов с учетом дополнительных атрибутов. Данный модуль использован в проекте «Антирутина-44» для выявления превышения цены закупки относительно нормальной рыночной цены.

Положения, выносимые на защиту. На защиту выносятся:

- алгоритм кластеризации базы государственных закупок;
- способ оценки рыночной цены товара;
- способ оценки типичности государственных закупок;
- способ оценки превышения рыночной цены;

Испытания на реальных данных. Предложенный в данной работе алгоритм был протестирован на реальных данных государственных закупок (более подробное описание см. в разделе «Вычислительный эксперимент»).

Обзор литературы Задача кластеризации базы государственных закупок представляет собой особый случай задачи кластеризации коротких текстов. Наиболее популярным типом подобных задач является задача кластеризации коротких тектовых сообщений в Twitter. В [1, 2] данная задача решается с использованием текстового моделирования. В [3, 4, 5] данная задача решается с помощью алгоритов, основанных на k-means. Рассмотрим подробнее эти и другие алгоритмы кластеризации.

Одним из наиболее популярных методов анализа и кластеризации текстов является тематическое моделирование. В простейшем виде задача тематического моделирования совдится к задаче разложения матрицы «слова-документы» на произведение матриц «слова-темы» и «темы-документы», где темы являются набором скрытых переменных, при этом их количество заранее задано. Однако, в таком виде задача может иметь несколько решений, кроме того, ее решение может быть плохо интерпертируемым. Для устранения проблем неединственности и неинтерпертируемости решения вводятся регуляризаторы. В [7, 9, 10] рассматриваются такие регуляризаторы, как разреженность матриц «слова-темы» и «темы-документы», сокращение числа тем, повышение различности тем и т.д. Наиболее полно задача тематического моделирования освещена в [6]. Кроме того, для построения тематических моделей с аддитивной регуляризацией для коллекций больших размеров существует библиотека BigARTM [8, 11].

K-means —один из старейших методов кластеризации. Данный алгоритм основан был представлен в [14], а в [15] был дополнен. Метод k-means основан на минимизации суммарного расстояния точек в кластере до центра кластера. K-means популярен

из-за своей простоты, однако, имеет несколько минусов: например, необходимость задания количества и определения начального расположения центров кластеров.

Еще одним известным способом кластеризации является иерархическая кластеризация. Ее принцип заключается в следующем: изначально каждая точка является кластером, затем кластеры постепенно объединяются. Для того, чтобы выбрать какие два кластера объединять в один, существует несколько методов, однако, наиболее популярны метод одиночной связи (метод ближайшего соседа) [19] и метод полной связи (метод дальнего соседа) [20]. Однако, оюба метода требуют точного подбора значения максимальног уровня иерархии, кроме того, зачастую, остается много шумовых точек, помеченных как отдельные кластеры.

Метод кластеризации, предложенный в данной работе, основан на алгоритме DBScan. Идея DBScan состоит в объединении в кластер плотно сгруппированных точек, попарно достижимых, и отбрасывании «шумовых» точек, расположенных в удалении от остальных [12]. Алгоритм DBScan определяет автоматически количество кластеров, является устойчивым к шуму и может находить кластеры любой формы, однако, плохо работает в случае, когда плотность точек в кластерах различна. Исправляет этот недостаток модификация HDBScan, иерархически подбирающая параметр пороговой плотности в кластере [13]. Однако, в HDBScan невозможно иерархически модифицировать функцию расстояния, что является главным отличием HDBScan от предложенной в данном исследовании модификации.

Алгоритм Spectral Custering основан на уменьшении размерности задачи при проецировании данных на собственные вектора матрицы попарных расстояний, соотвествующие k наибольшим по модулю собственным значениям. После этого применяется алгоритм, схожий с k-means [17, 18].

Алгоритм Affinity propagation основан на идее «передачи сообщений» между объектами выборки и оценивании того, насколько один объект может служить центром или представителем для другого [16].

Эти и другие алгоритмы кластеризации наиболее широко освещены в [21].

Также в задаче кластеризации текстов возникает подзадача предобработки данных. Библиотеки pymorphy2 и nltk предоставляют широкий функционал для обработки слов в текстах и имеют документацию, в которой подробно описаны методы подобной обработки [23, 22].

2 Постановка задачи

Имеется база государственных закупок $\mathcal{D} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$. Каждая закупка $\mathbf{x}$ характеризуется рядом признаков:

- x^{text} —текстовое описание закупки (предлагается лемматизировать тектовое описание и удалить из него стоп-слова, а также все числа и конструкции типа «измерение+единица измерения»);
- x^{meas} —синтетический признак «измерения закупки» (выделенные и удаленные из текстового описания конструкции типа «измерение+единица измерения»);
- x^{price} —цена закупки;
- x^{count} —количество товара в закупке;
- x^{date} —дата совершения закупки;
- x^{reg} —регион совершения закупки;
- x^{unit} —единица товара закупки;
- x^{RCPEA} —код ОКПД (Общероссийский классификатор продукции по видам экономической деятельности) закупки.

Требуется:

- оценить типичность госзакупки $\mathbf{x} \in \mathcal{D}$;
- оценить рыночную цену товара, представленного в закупке;
- оценить превышение рыночной цены.

Предлагается модель:

$$\text{typicality}(\mathbf{x}) = \begin{cases} 1, \text{ если } x^{\text{price}} \leq \text{MarketPrice}; \\ 0, \text{ иначе}, \end{cases}$$

$$\text{excess}(\mathbf{x}) = (x^{\text{price}} \; - - - \text{MarketPrice})_+.$$

Для получения оценок рыночной цены товара предлагается кластеризовать базу государственных закупок. Требуется кластеризовать базу закупок:

$$a : \mathcal{D} \to \mathcal{C}, \qquad \mathcal{C} = \{C_1, \ldots, C_K\}.$$

Предполагается, что цены госзакупок из одного кластера имеют гамма-распределение:

$$p(x^{\text{price}}) = \Gamma(k_C, \lambda_C), \qquad a(\mathbf{x}) = C.$$

Тогда цены всех госзакупок образуют следующую смесь гамма-распределений:

$$p(x^{\text{price}}) = \sum_{C \in \mathcal{C}} [a(\mathbf{x}) = C] \Gamma(k_C, \lambda_C).$$

Экспертное оценка качества кластера:

$$g(C) = \left[\frac{\sigma_C}{\mu_C} \leq \theta^{\text{price}} \right] [|C| \geq \text{MinPts}] \left[\max_{\substack{\mathbf{x}_1, \mathbf{x}_2: \\ a(\mathbf{x}_1) = a(\mathbf{x}_2) = C}} \text{dist}(\mathbf{x}_1, \mathbf{x}_2) \leq \theta^{\text{dist}} \right],$$

где μ_C, σ_C —средняя цена и стандартное отклонение цены в C, $|C| = |\mathbf{x} : a(\mathbf{x}) = C|$, $\text{MarketPrice}(C) = \mu_C + \sigma_C$ —рыночная цена в C, $\theta^{\text{price}}, \text{MinPts}, \theta^{\text{dist}}$ —задаваемые экспертом пороги маскимального разброса цен в кластере, минимального размера кластера и максимально возможного диаметра кластера, соответственно.

Также договоримся, что в дальнейшем запись $\mathbf{x} \in C$ будет равнозначна записи $a(\mathbf{x}) = C$.

Тогда задача кластеризации базы закупок формулируется как следующая задача экспертно-интерпертуемой декомпозиции смеси распределений (задача многокритериальной оптимизации):

$$\begin{cases} L(\mathcal{D}, \mathcal{C}) = -\frac{1}{|\mathcal{D}|} \sum\limits_{\mathbf{x} \in \mathcal{D}} \ln \sum\limits_{C \in \mathcal{C}} [\mathbf{x} \in C] \Gamma(k_C, \lambda_C)|_{\mathbf{x}} \to \min\limits_{a: \mathcal{D} \to \mathcal{C}}, \\ \frac{\sum\limits_{C \in \mathcal{C}} g(C)}{|\mathcal{C}|} \to \max\limits_{a: \mathcal{D} \to \mathcal{C}}. \end{cases}$$

Для упрощения выбора оптимального алгоритма кластеризации пердлагается также рассмотреть следующий вариант постановки задачи:

$$\begin{cases} \frac{\sum\limits_{C \in \mathcal{C}} g(C)}{|\mathcal{C}|} \to \max\limits_{a: \mathcal{D} \to \mathcal{C}}, \\ L(\mathcal{D}, \mathcal{C}) = -\frac{1}{|\mathcal{D}|} \sum\limits_{\mathbf{x} \in \mathcal{D}} \ln \sum\limits_{C \in \mathcal{C}} [\mathbf{x} \in C] \Gamma(k_C, \lambda_C)|_{\mathbf{x}} \leq \theta^{\text{expert}} \end{cases}$$

3 Кластеризация базы государственных закупок

3.1 Алгоритм решения поставленной задачи

Для уменьшения размерности задачи и разбиения исходной базы на набор выборок предлагается использовать тематическое моделирование с аддитивной регуляризацией. Результатом работы алгоритма тематического моделирования будет набор

тем $\mathcal{T} = \{T_1, \ldots, T_N\}$ Каждая полученная тема T_i будет соответствовать одной выборке.

Для выделения различных форм товаров и определения типичности закупки предлагается кластеризовать все закупки с помощью иерархической модификации алгоритма DBScan. Таким образом, алгоритм Hierarchical DBScan будет искомым алгоритмом a (см. раздел «Постановка задачи»). В разделе «Вычислительный эксперимент» будет показано, что алгоритм Hierarchical DBScan действительно наилучшим образом удовлетворяет поставленой задаче. Полученные в результате применения алгоритма Hierarchical DBScan кластеры будут соответствовать одной форме выпуска товара.

Алгоритм решения задачи представлен в Алг. 3.1.

Далее подробнее опишем алгоритм тематического моделирования и иерархическую модификацию алгоритма DBScan.

3.2 Тематическое моделирования

Исходная база данных закупок представляется в виде коллекции текстовых описаний закупок $\mathcal{D}$. Для данной коллекции формируется словарь W используемых в коллекции слов. Каждая закупка $\mathbf{x} \in \mathcal{D}$ (здесь и далее в этом подразделе под закупкой понимается ее текстовое описание и обозначать как x) представляет собой набор слов из W длиной n_x.

Предлагается не использовать в качестве слов конструкции типа «измерение+единица измерения», например, «100 мг».

Предполагается, что существует конечное множество $\mathcal{T}$ скрытых тем. Употребление слова w в каждой закупке x связано с некоторой темой $T \in \mathcal{T}$, которая не известна. Вводятся следующие гипотезы:

- база закупок —набор троек (x, w, T), каждая из троек случайна и независима от остальных троек;
- появление слова w в закупке x не зависит от документа и зависит только от тематической принадлежности T закупки x:

$$p(w|T, x) = p(w|T),$$

$$p(x|T, w) = p(x|T),$$

$$p(x, w|T) = p(x|T)p(w|T);$$

Алгоритм 3.1. Алгоритм решения поставленной задачи

Вход: $\mathcal{D}$ —база данных закупок, ε, MinPts —параметры алгоритма DBScan;

Выход: $\mathcal{C}$ —разбиение выборки $\mathcal{D}$ на кластеры; $\{\text{typicality}(\mathbf{x}), \text{excess}(\mathbf{x})\}$ — типичность и превышение для каждой закупки $\mathbf{x} \in \mathcal{D}$

$\mathcal{C} := \varnothing$;
$\mathcal{T} := \text{TopicModeling}(\mathcal{D})$;
$\{\mathcal{X}_1, \ldots, \mathcal{X}_{|\mathcal{T}|}\}$ —набор выборок;
для всех $\mathbf{x} \in \mathcal{D}$
    $\tilde{T} := \arg\max\limits_{T \in \mathcal{T}} p(T|x)$;
    $X_{\tilde{T}}.\text{append}(\mathbf{x})$;
для всех X_T
    $\mathcal{C}_T := \text{HierarchicalDBScan}(X_T)$ —набор кластеров $\{C_{T1}, \ldots, C_{TN_T}\}$ в данной выборке;
    $\mathcal{C}.\text{append}(\mathcal{C}_T)$;
для всех $C \in \mathcal{C}$
    $\mu_C := \frac{\sum\limits_{\mathbf{x} \in C} x^{\text{price}} * x^{\text{count}}}{\sum\limits_{\mathbf{x} \in C} x^{\text{count}}}$;
    $\sigma_C := \sqrt{\frac{\sum\limits_{\mathbf{x} \in C} x^{\text{count}} * (x^{\text{price}} - \mu_C)^2}{\sum\limits_{\mathbf{x} \in C} x^{\text{count}}}}$;
    для всех $\mathbf{x} \in C$
        если $x^{\text{price}} \leq \mu_C + \sigma_C$ **то**
            $\text{typicality}(\mathbf{x}) := 1$;
            $\text{excess}(\mathbf{x}) := 0$;
        иначе
            $\text{typicality}(\mathbf{x}) := 0$;
            $\text{excess}(\mathbf{x}) := x^{\text{price}} \ - - - \ \mu_C - \sigma_C$;

- как следствие, вероятность появления слова w в закупке x выражается следующей формулой:

$$p(w|x) = \sum_{T \in \mathcal{T}} p(w|T) p(T|x);$$

- количество тем $|\mathcal{T}|$ много меньше размера базы закупок $|\mathcal{D}|$ и количества слов $|W|$.

Введем также следующие величины:

n_{xw} —количество вхождений слова w в закупку x,

$$n_x = \sum_{w \in W} n_{xw} \text{ —длина закупки } x,$$

$$\Phi = [p(w|T)]_{|W| \times |\mathcal{T}|},$$

$$\Theta = [p(T|x)]_{|\mathcal{T}| \times |\mathcal{D}|}$$

Тогда задача тематического моделирования,используя принцип максимума правдоподобия, формулируется следующим образом:

$$p(\mathcal{D}, \Phi, \Theta) = Const * \prod_{x \in \mathcal{D}} \prod_{w \in x} p(x, w)^{n_{xw}} \to \max_{\Phi, \Theta}.$$

Логарифмируя и избавляясь от постоянного слагаемого, получаем следующую формулировку задачи тематического моделирования:

$$L(\Phi, \Theta) = \sum_{x \in \mathcal{D}} \sum_{w \in x} n_{xw} \ln \sum_{T \in \mathcal{T}} p(w|T) p(T|x) \to \max_{\Phi, \Theta}.$$

Введем дополнительные величины:

$$F = [\hat{p}(w|x)]_{|W| \times |\mathcal{D}|} = \left[\frac{n_{xw}}{n_x} \right]_{|W| \times |\mathcal{D}|},$$

$\hat{p}_{wx} = \hat{p}(w|x) = \dfrac{n_{xw}}{n_x}$ – оценка вероятности слова w входить в заупку x,

$H_{xwT} = p(T|x, w) = \dfrac{p(w|T)p(T|x)}{p(w|x)} = \dfrac{\varphi_{wT}\theta_{Tx}}{\sum\limits_{T' \in \mathcal{T}} \varphi_{wT'}\theta_{T'x}}$ —вероятность того,

что тема T описывает вхождения слова w в закупку x,

$\hat{n}_{xwT} = n_{xw} H_{xwT}$ —оценка числа вхождений слова w в закупку x, связанных с темой T,

$\hat{n}_{wT} = \sum\limits_{x \in \mathcal{D}} n_{xw} H_{xwT}$ —оценка числа троек (x, w, T), в которых слово из x связано с темой T,

$\hat{n}_{xT} = \sum\limits_{w \in W} n_{xw} H_{xwT}$ — оценка числа троек (x, w, T), в которых слово w связано с темой T,

$\hat{n}_T = \sum\limits_{w \in W} \hat{n}_{wT}$ —оценка числа троек (x, w, T), связанных с темой T,

$\varphi_{wT} = \hat{p}(w|T) = \dfrac{\hat{n}_{wT}}{\hat{n}_T}$ —оценка вероятности того, что появление слова w связано с темой T,

$\theta_{Tx} = \hat{p}(T|x) = \dfrac{\hat{n}_{xT}}{n_x}$ —оценка вероятности того, что закупка x описывается темой T.

Заметим, что, используя введенные обозначенияб задачу тематического моделирования можно также сформулировать следующим образом:

$$||F - \Phi\Theta||^2 \to \min_{\Phi, \Theta}$$

Алгоритм 3.2. EM-алгоритм для решения задачи тематического моделирования

Вход: $\mathcal{D}$ —коллекция текстовых описаний закупок,W —словарь, $[n_{xw}]_{|W|\times|\mathcal{D}|}, n_x$ для каждого $x \in \mathcal{D}$, $|\mathcal{T}|$ —количество тем, θ^{iter} —максимальное количество итераций алгоритма, θ^{lkhd} —порог сходимости правдоподобия.

Выход: Φ, Θ.

1: инициализируем Φ, Θ;
2: $L(\Phi,\Theta) := \sum\limits_{x\in\mathcal{D}}\sum\limits_{w\in x} n_{xw} \ln \sum\limits_{T\in\mathcal{T}} \varphi_{wT}\theta_{Tx}$;
3: iter $:= 0$;
4: $\Delta L(\Phi,\Theta) := \theta^{\text{lkhd}} + 1$;
5: **пока** $(\Delta L(\Phi,\Theta) > \theta^{\text{lkhd}})$ $\quad OR \quad$ $(\text{iter} < \theta^{\text{iter}})$
6: **E-шаг**: $H_{xwT} := p(T|x,w) = \frac{\varphi_{wT}\theta_{Tx}}{\sum\limits_{T'\in\mathcal{T}} \varphi_{wT'}\theta_{T'x}}$
7: **M-шаг**: $\hat{n}_{wT} := \sum\limits_{x\in\mathcal{D}} n_{xw}H_{xwT}$
8: **M-шаг**: $\hat{n}_{xT} := \sum\limits_{w\in W} n_{xw}H_{xwT}$;
9: **M-шаг**: $\hat{n}_{T} := \sum\limits_{w\in W} \hat{n}_{wT}$;
10: **M-шаг**: $\varphi_{wT} := \frac{\hat{n}_{wT}}{\hat{n}_T}$;
11: **M-шаг**: $\theta_{Tx} := \frac{\hat{n}_{xT}}{n_x}$;
12: $L_{old}(\Phi,\Theta) := L(\Phi,\Theta)$;
13: $L(\Phi,\Theta) := \sum\limits_{x\in\mathcal{D}}\sum\limits_{w\in x} n_{xw} \ln \sum\limits_{T\in\mathcal{T}} \varphi_{wT}\theta_{Tx}$;
14: $\Delta L(\Phi,\Theta) := |L(\Phi,\Theta) \ --- \ L_{old}(\Phi,\Theta)|$;
15: iter $:=$ iter $+ 1$

Предлагается решать задачу тематического моделирования с помощью EM-алгоритма (см. Алг. 3.2): Заметим, что найденное с помощью данного алгоритма решение в общем случае неединственно в силу следующего рассуждения:

$$\Phi\Theta = (\Phi S^{-1})(S\Theta) = \Phi'\Theta'.$$

Необходимо только, чтобы матрицы Φ' и Θ были стохастическими.

Для повышения устойчивости решения (избавления от неединственности решения) и повышения интерпретируемости тем предлагается использовать регуляризацию. В этом случае задача тематического моделирования переформулируется следующим образом:

$$L(\Phi,\Theta) + R(\Phi,\Theta) \to \max_{\Phi,\Theta},$$

$$R(\Phi,\Theta) = \sum_{i=1}^{n} \tau_i R_i(\Phi,\Theta),$$

$R_i(\Phi, \Theta)$ —дифференцируемые по $\varphi_{wT}, \theta_{Tx}$ функции,

$$\tau_i \geq 0 \ \forall i = 1, \ldots, n.$$

Тогда формулы пересчета $\varphi_{wT}, \theta_{Tx}$ на M-шаге алгоритма 3.2 изменятся следующим образом:

$$\varphi_{wT} = \frac{(\hat{n}_{wT} + \varphi_{wT}\frac{\partial R(\Phi,\Theta)}{\varphi_{wT}})_+}{\sum\limits_{u \in W}(\hat{n}_{uT} + \varphi_{uT}\frac{\partial R(\Phi,\Theta)}{\varphi_{uT}})_+},$$

$$\theta_{Tx} = \frac{(\hat{n}_{xT} + \theta_{Tx}\frac{\partial R(\Phi,\Theta)}{\theta_{Tx}})_+}{\sum\limits_{T' \in \mathcal{T}}((\hat{n}_{xT'} + \theta_{T'x}\frac{\partial R(\Phi,\Theta)}{\theta_{T'x}})_+}.$$

Предлагается использовать следующие регуляризаторы:

- разреживание матрицы Φ

$$R_1(\Phi, \Theta) = -\sum_{T \in \mathcal{T}}\sum_{w \in W}\varphi_{wT};$$

- разреживание матрицы Θ

$$R_2(\Phi, \Theta) = -\sum_{x \in \mathcal{D}}\sum_{T \in \mathcal{T}}\theta_{Tx};$$

- повышение различности тем

$$R_3(\Phi, \Theta) = -\frac{1}{2}\sum_{T \in \mathcal{T}}\sum_{T' \in \mathcal{T} \setminus T} cov(\varphi_T, \varphi_{T'}),$$

$$cov(\varphi_T, \varphi_{T'}) = \sum_{w \in W}\varphi_{wT}\varphi_{wT'};$$

- сокращение числа тем

$$R_4(\Phi, \Theta) = -\sum_{T \in \mathcal{T}}\ln\sum_{x \in \mathcal{D}}\frac{n_x}{\hat{n}_T}\theta_{Tx}.$$

Оценкой качества тематического моделирования служат критерии, перечисленные ниже.

- Перплексия:

$$\mathcal{P}(\mathcal{D}) = \exp\left(-\frac{1}{n}L(\Phi, \Theta)\right) = \exp\left(-\frac{1}{n}\sum_{x \in \mathcal{D}}\sum_{w \in x} n_{xw}\ln\sum_{T \in \mathcal{T}} p(w|T)p(T|x)\right).$$

- Разреженность матрицы Θ «темы-документы»:

$$S_\Theta = \frac{\sum\limits_{T\in\mathcal{T}}\sum\limits_{x\in\mathcal{D}}[p(T|x)=0]}{|\mathcal{T}||\mathcal{D}|}.$$

- Разреженность матрицы Φ «слова-темы»:

$$S_\Phi = \frac{\sum\limits_{w\in W}\sum\limits_{T\in\mathcal{T}}[p(w|T)=0]}{|W||\mathcal{T}|}.$$

Предполагается, что в результате работы тематического моделирования исходная база закупок разобьется на набор непересекающихся выборок, соответсвующих темам из $\mathcal{T}$ и представляющих собой набор товаров. Будем относить к выборке X_T , соответствующей теме T, закупку x, если $T = \arg\max\limits_{T'\in\mathcal{T}} p(T'|x)$.
Данный этап позволяет уменьшить размерность решаемой задачи и сформировать более однородные, чем исходная база закупок, выборки для последующей кластеризации.

3.3 Базовый алгоритм DBScan

Прежде чем описать иерархическую модификацию алгоритма DBScan, используемую для решения поставленной задачи, опишем базовый алгоритм DBScan.
Будем называть точку P корневой, если на расстоянии не более ε от нее находится не менее MinPts точек. Данные точки назовем непосредственно плотно-достижимыми из P. Ни одна точка не является непосредственно плотно-достижимой из некорневой точки.
Будем называть точку Q плотно-достижимой из точки P, если существует путь $P_1, \ldots, P_n$, удовлетворяющий следующим условиям: $P_1 = P$, $P_n = Q$, P_{i+1} непосредственно плотно-достижима из P_i для всех $i = 1, \ldots, n-1$.
Будем называть точки, которые не являются плотно-достижимыми ни из одной другой точки выбросами.
Будем называть точки P и Q плотно-соединенными, если существует точка O такая, что P и Q плотно-достижимы из O.
Тогда кластер будет формироваться как набор всех плотно-достижимых точек из заданной корневой точки P. Заметим, что один и тот же кластер может быть сформирован на основе различных корневых точек.
Каждый кластер удовлетворяет следующим свойствам:

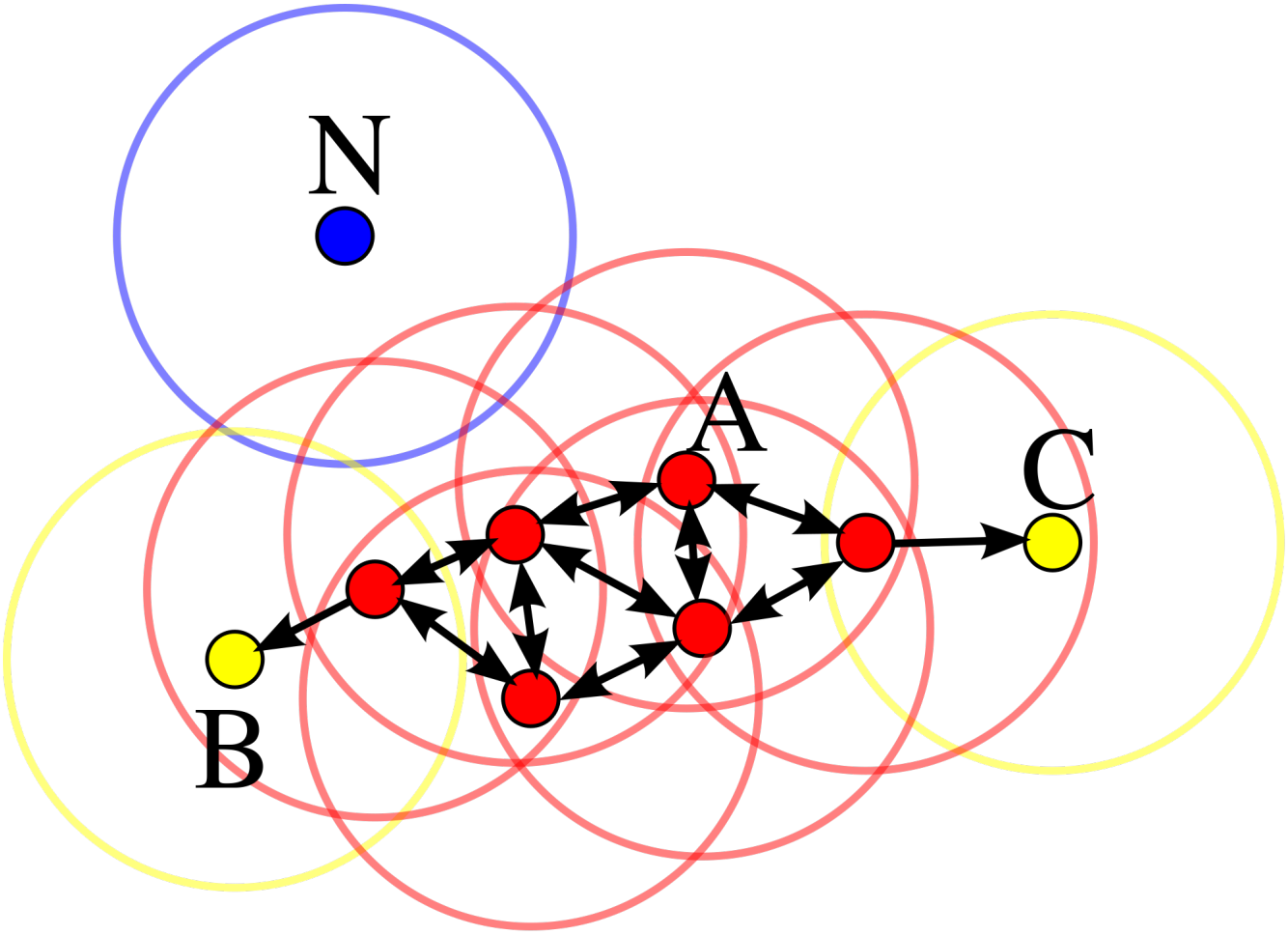

Рис. 1: Пример классификации точек: красные точки —корневые, желтые —плотно-достижимые из корневых, синие —выбросы (MinPts = 3)

- все точки внутри кластера являются попарно плотно-соединенными;

- если точка P плотно-достижима из некоторой точки Q кластера, то P также принадлежит данному кластеру.

Пример классификации точек представлен на рис. 1 (рисунок взят из [24]). На данном рисунке красными точками отмечены корневые, желтымые —плотно-достижимые из корневых, синими —выбросы. Значения параметра MinPts = 3 , параметр ε равен радиусу отмеченных окружностей. Также этот рисунок иллюстрирует работу алгоритма DBScan —корневые и плотно-достижимые из корневых точки образуют кластер, кроме того, имеется один выброс.

Алгоритм 3.3 представляет собой описание алгоритма кластеризации DBScan и ссылается на функции 3.4, 3.5.

Алгоритм 3.3. Алгоритм DBScan

Вход: $\mathcal{X}$ —выборка, ε, MinPts —параметры алгоритма DBScan;

Выход: $\mathcal{C}$ —разбиение выборки $\mathcal{X}$ на кластеры;

$\mathcal{C} := \varnothing$
$i := 0$;
$C_i := \varnothing$;
$C_{\text{noise}} := \varnothing$;
Visited $= \varnothing$;
NotVisited $= \mathcal{X}$;
пока NotVisited $\neq \varnothing$
    $P :=$ NotVisited.pop;
    Visited.append(P);
    NeighborPts := regionQuery($P, \mathcal{X}, \varepsilon$);
    если |NeighborPts| $<$ MinPts **то**
        $C_{\text{noise}} := C_{\text{noise}}$.append($P$);
    иначе
        $C_i, C_{\text{noise}} :=$ expandCluster(P, NeighborPts, $C_i, C_{\text{noise}}, \mathcal{C}, \mathcal{X}, \varepsilon$, MinPts, Visited, NotVisited);
        $\mathcal{C}$.append(C_i);
        $i := i + 1$;
        $C_i = \varnothing$;
$\mathcal{C}$.append(C_{noise});

3.4 Используемая функция расстояния между закупками

Прежде чем описать иерархическую модификацию алгоритма DBScan, используемую для решения поставленной задачи, опишем используемую функцию расстояния между закупками.

Предлагается для определения расстояния между закупками использовать нижеперечисленные признаки закупок:

- текстовое описание закупки x^{text} (предлагается лемматизировать тектовое описание и удалить из него стоп-слова, а также все числа и конструкции типа «измерение+единица измерения»);
- синтетический атрибут «измерения закупки» x^{meas} (выделенные и удаленные из текстового описания конструкции типа «измерение+единица измерения»);

Алгоритм 3.4. Функция expandCluster

Вход: $P, \text{NeighborPts}, C_i, C_{\text{noise}}, \mathcal{C}, \mathcal{X}, \varepsilon, \text{MinPts}, \text{Visited}, \text{NotVisited}$;

Выход: C_i, C_{noise};

```
1: C_i.append(P);
2: для всех P' ∈ NeighborPts
3:    если P' ∈ NotVisited то
4:       NotVisited.delete(P');
5:       Visited.append(P');
6:       NeighborPts' := regionQuery(P', 𝒳, ε);
7:       если |NeighborPts'| ≥ MinPts то
8:          NeighborPts.join(NeighborPts');
9:    если P' ∈ C_noise то
10:      C_noise.delete(P');
11:      C_i.appendP';
12:   иначе если P' ∉ C, ∀C ∈ 𝒞 то
13:      C_i.append(P')
```

Алгоритм 3.5. Функция regionQuery

Вход: $P, \mathcal{X}, \varepsilon$;

Выход: NeighborPts;

```
1: NeighborPts = ∅;
2: для всех P' ∈ 𝒳
3:    если dist(P, P') ≤ ε то
4:       NeighborPts.append(P')
```

- цена закупки x^{price}.

Определим следующие функции близости между двумя закупками:

- вложенность тектового описания одной закупки в другую:

$$\text{sim}_1(\mathbf{x}_1, \mathbf{x}_2) = \frac{|\text{set}(x_1^{\text{text}}) \cap \text{set}(x_2^{\text{text}})|}{\min(|\text{set}(x_1^{\text{text}})|, |\text{set}(x_2^{\text{text}})|)};$$

- близость по мере Жаккара измерений закупок:

$$\text{sim}_2(\mathbf{x}_1, \mathbf{x}_2) = \frac{|\text{set}(x_1^{\text{meas}}) \cap \text{set}(x_2^{\text{meas}})|}{|\text{set}(x_1^{\text{meas}}) \cup \text{set}(x_2^{\text{meas}})|};$$

- близость цен закупок:

$$\mathrm{sim}_3(\mathbf{x}_1, \mathbf{x}_2) = 1 - \frac{|x_1^{\mathrm{price}} - x_2^{\mathrm{price}}|}{x_1^{\mathrm{price}} + x_2^{\mathrm{price}}}.$$

Тогда функцию близости двух закупок определим следующим образом:

$$\mathrm{sim}(\mathbf{x}_1, \mathbf{x}_2) = \omega_1 \mathrm{sim}_1(\mathbf{x}_1, \mathbf{x}_2) + \omega_2 \mathrm{sim}_2(\mathbf{x}_1, \mathbf{x}_2) + \omega_3 \mathrm{sim}_3(\mathbf{x}_1, \mathbf{x}_2),$$

$$\omega_1 + \omega_2 + \omega_3 = 1.$$

Соответственно, расстояние между двумя закупками определим следующим образом:

$$\mathrm{dist}(\mathbf{x}_1, \mathbf{x}_2) = 1 \ - - - \mathrm{sim}(\mathbf{x}_1, \mathbf{x}_2).$$

3.5 Алгоритма Hierarchical DBScan

Для решения поставленной задачи предлагается использовать иерахическую модификацию алгоритма DBScan.

Предлагается на первой итерации (нулевом уровне иерархии) кластеризовать всю данную выборку. Затем, изменяя параметры алгоритма и веса признаков в метрике, предалагается на следующем уровне иерархии кластеры, полученные на предыдущем уровне иерархии разбивать на новые кластеры. Таким образом, на ранних этапах произойдет разбиение выборки на кластеры, соответстующие различным товарам, а на поздних этапах в кластеры будут входить близкие по цене закупки, соответствующие одной форме товара. Алгоритм 3.6 представляет собой описание алгоритма Hierarchial DBScan с изменюящейся в зависимости от уровня иерархии функцией расстояния.

Описание эвристик для выбора $\varepsilon, \omega_1, \omega_2, \omega_3$ будет приведено в следующем подразделе.

Для уменьшения времени работы алгоритма предлагается следующая модификация: предлагается не разбивать на кластеры следующего уровня те кластеры, которые удовлетворяют экспертному критерию качества кластера (см. раздел «Постановка задачи»). Экспертно-модифицированный алгоритм Hierarchical DBScan представлен в 3.7.

3.6 Эвристики для выбора параметров алгоритма

Точно подобрать оптимальные параметры алгоритма и веса признаков в функции расстояния представляется трудновыполнимой задачей, однако предложены

Алгоритм 3.6. Hierarchical DBScan

Вход: $\mathcal{X}$ —выборка, MinPts —параметр алгоритма DBScan, θ^{lvl} —максимальный уровень иерархии;

Выход: $\mathcal{C}$ —разбиение выборки $\mathcal{X}$ на кластеры;

1: $\mathcal{C} := \varnothing$
2: LOC := $[\{\mathcal{X}, 0\}]$;
3: **пока** LOC $\neq \emptyset$
4: $\{D, \text{level}\} :=$ LOC.pop;
5: dist = dist$(\omega_1(\text{level}), \omega_2(\text{level}), \omega_3(\text{level}))$ – меняем функцию расстояния;
6: $\varepsilon :=$ epsilonSelection$(\text{MinPts}, D, \text{level})$;
7: $\tilde{\mathcal{C}} :=$ DBScan$(D, \varepsilon, \text{MinPts})$
8: **если** level $< \theta^{\text{lvl}}$ **то**
9: **для всех** $C \in \tilde{\mathcal{C}}$
10: LOC.append$(\{C, \text{level} + 1\})$
11: **иначе**
12: **для всех** $C \in \tilde{\mathcal{C}}$
13: $\mathcal{C}$.append(C)

некоторые эвристики для подбора данных параметров. Пусть $k-\text{Nbr}(\mathbf{x})$ —множество, состоящее из k ближайших соседей точки $\mathbf{x}$, тогда в качестве оценки параметра ε предлагается брать следующее значение:

$$\varepsilon = \frac{1}{|\mathcal{X}|} \sum_{fatx \in \mathcal{X}} \frac{\sum\limits_{\mathbf{x}' \in \text{MinPts}-\text{Nbr}(\mathbf{x})} dist(\mathbf{x}, \mathbf{x}')}{\text{MinPts}}.$$

Параметр $MinPts$ выбирается экспертно и интерпретируется как минимальное количество точек в кластере.

Также предлагается следующая эвристика для подбора значения параметра ε:

$$\varepsilon = \varepsilon_0 q^{-\text{level}}.$$

Подбор значений $\omega_1, \omega_2, \omega_3$ осуществляется из следующих соображений: на ранних этапах кластеризации предполагается выборку разделить на кластеры соответсвующие определенным товарам, т.е наибольшее значение принимает ω_1. В дальнейшем кластеры интерпретируются как форма выпуска товара и наибольшее значение принимает ω_2. На последних этапах следует увеличить значение ω_3 —предполагается, что

Алгоритм 3.7. Экспертно-модифицированный алгоритм Hierarchical DBScan

Вход: $\mathcal{X}$ —выборка, MinPts —параметр алгоритма DBScan, θ^{lvl} —максимальный уровень иерархии;

Выход: $\mathcal{C}$ —разбиение выборки $\mathcal{X}$ на кластеры;

$\mathcal{C} := \varnothing$
LOC := $[\{\mathcal{X}, 0\}]$;
пока LOC $\neq \emptyset$
  $\{D, \text{level}\} :=$ LOC.pop;
  dist = dist$(\omega_1(\text{level}), \omega_2(\text{level}), \omega_3(\text{level}))$ – меняем функцию расстояния;
  $\varepsilon :=$ epsilonSelection(MinPts, D, level);
  $\tilde{\mathcal{C}} :=$ DBScan$(D, \varepsilon, \text{MinPts})$
  для всех $C \in \tilde{\mathcal{C}}$
    если $(\text{level} < \theta^{\text{lvl}})$ AND $(g(C) == 0)$ **то**
      LOC.append$(\{C, \text{level} + 1\})$
    иначе
      $\mathcal{C}$.append(C);

закупки одной формы выпуска будут иметь схожую цену, но цена из-за инфляции может различаться в зависимости от даты совершения закупки.

4 Вычислительный эксперимент

4.1 Описание данных

Описание выборки Для вычислительного эксперимента использовались закупки, имеющие один из перечисленных ниже кодов ОКПД:

- 33.10.15.131: Инструменты режущие и ударные с острой (режущей) кромкой однолезвийные;
- 21.20.10.239: Препараты для лечения нервной системы;
- 24.42.13.779: Средства противораковые;
- 24.42.13.796: Средства противовирусные;
- 36.63.21.110: Ручки шариковые;

- 33.10.15.121: Шприцы-инъекторы медицинские многоразового и одноразового использования с инъекционными иглами и без них;
- 30.01.24.110: Части и принадлежности копировально-множительных машин.

Размер полученной таким образов выборки —464258 закупки (суммарная стоимость закупок более 125 млрд. рублей), при этом количество уникальных слов (после лематизации, о которой подробнее будет написано ниже), встречающихся не менее 5 раз —8329 слов.

Лемматизация Для построения тематических моделей и признакового пространства для кластеризации была проведена процедура *лемматезиции* слов. *Лемматизация* —это приведение слов в тексте к их нормальной (начальной) форме. Данная процедура позволяет сократить словарь (размерность признакового пространства). В русском языке начальными формами считаются:

- для существительных —именительный падеж, единственное число;
- для прилагательных —именительный падеж, единственное число, мужской род;
- для количественных числительных —именительный падеж;
- для порядковых числительных —именительный падеж, единственное число, мужской род;
- для глаголов —глагол в инфинитиве;
- для большинства местоимений-существительных —именительный падеж, единственное число;
- для местоимений-прилагательных —именительный падеж, единственное число, мужской род;
- для местоимений-числительных —именительный падеж;
- для причастий —глагол в инфинитиве;
- для деепричастий —глагол в инфинитиве.

Для лемматизации использовалась библиотека pymorphy2 [23].

Выделение стоп-слов. Для построения тематических моделей и признакового пространства для кластеризации была проведена процедура выделения стоп-слов. Стоп-слова —это слова, которые встречаются в текстах различной тематики и не несут смысловой нагрузки. Как правило, к стоп-словам относят предлоги, союзы, междометия, частицы и т.п.

Для выделения стоп-слов использовалась библиотек nltk [22], содержащая приведенный ниже список стоп-слов (см. Табл. 1).

Таблица 1: Список стоп-слов в библиотеке nltk

и	в	во	не	что	он	на	я	с	никогда
то	все	она	так	его	но	да	ты	к	сейчас
за	бы	по	только	ее	мне	было	вот	от	всех
о	из	ему	теперь	когда	даже	ну	вдруг	ли	были
ни	быть	был	него	до	вас	нибудь	опять	уж	этого
потом	себя	ничего	ей	может	они	тут	где	есть	чего
мы	тебя	их	чем	была	сам	чтоб	без	будто	ней
себе	под	будет	ж	тогда	кто	этот	того	потому	раз
ним	здесь	этом	один	почти	мой	тем	чтобы	нее	больше
а	вы	нет	или	для	тоже	совсем	куда	зачем	нельзя
можно	при	наконец	два	об	другой	хоть	после	над	как
через	эти	нас	про	всего	них	какая	много	разве	со
моя	впрочем	хорошо	свою	этой	перед	иногда	лучше	чуть	том
такой	им	более	всегда	конечно	всю	между	тот	три	эту
у	же	меня	еще	если	уже	вам	ведь	надо	какой

Выделение синтетического признака «измерения закупки» Для выделения синтетического признака «измерения закупки» из текстового описания (до леммматизации и удаления стоп-слов) предлагается воспользоваться следующим регулярным выражением:

$$([№|n|][-]^*|[0-9]^+[,.]^?[0-9]^*)|([0-9]^+[,.]^?[0-9]^*[\ ]^*[-|a-z|\%][-|a-z|/|\%|0-9]^*).$$

Здесь a^* означает, что подвыражение a встречается любое количество раз подряд; a^+ означает, что подвыражение a встречается любое большее, чем 0, количество раз подряд; $a^?$ означает, что подвыражение a встречается 0 или 1 раз подряд.

Все совпадения с данным выражением переносятся в синтетический признак x^{meas} «измерения закупки» и удаляются из текстового описания. После этого текстовой описание проходит процедуры лемматизации и удаления стоп-слов.

4.2 Вычислительный эксперимент и анализ его результатов

В первой части вычислительного эксперитмента было построено 4 тематические модели:

- PLSA —тематическая модель без регуляризаторов;
- ARTM —тематическая модель с регуляризаторами разреженности матриц Φ и Θ и регуляризатором сокращения числа тем;
- ARTM1 —тематическая модель с регуляризаторами разреженности матриц Φ и Θ и регуляризатором повышения различности тем;
- ARTM2 —тематическая модель с регуляризаторами разреженности матриц Φ и Θ.

При построении моделей количество тем было выбрано равным 100. Модель ARTM1 сократила это количество до 18.
Тематические модели были построены с использованием библиотеки BigARTM [11].
Графики 2 и 3 показывают сравнение перплексии моделей с регуляризацией с перплексией модели PLSA. Как видно из этих графиков, падение перплексии для моделей ARTM и ARTM2 сравнительно невелико (для модели ARTM несколько больше, чем для модели ARTM2). Для модели ARTM1 ухудшение перплексии достаточно велико. Таким образом, модели PLSA, ARTM и ARTM1 лучше предсказывают появление терминов $w \in W$ в текстах закупок $x \in \mathcal{D}$.

Графики 4 и 5 показывают сравнение разреженности S_Θ матрицы Θ «темы-документы» для моделей с регуляризацией с разреженностью S_Θ для модели PLSA. Как видно из этих графиков, наибольшая разреженность достигается в модели ARTM. Разреженность, достигающаяся в моделях ARTM и ARTM1 достаточна для однозночного отнесения большинства закупок к определенной теме, что позволяет сформировать набор подвыборок для последующих экспериментов. Разреженность,

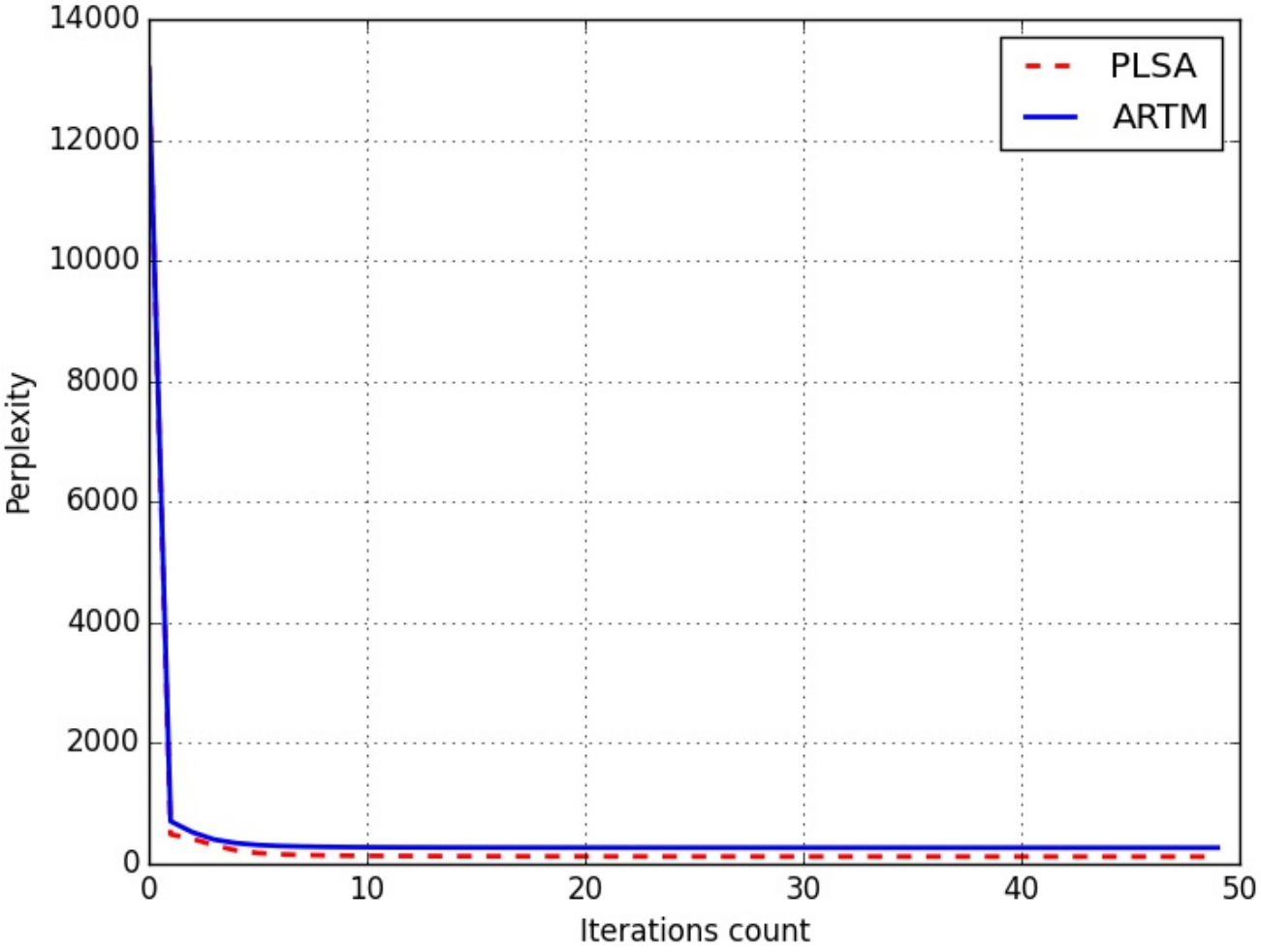

Рис. 2: Сравнение перплексии для моделей PLSA и ARTM

достигаемая в моделях PLSA и ARTM2 недостаточна для этих целей.

Графики 6 и 7 показывают сравнение разреженности S_Φ матрицы Φ «слова-темы» для моделей с регуляризацией с разреженностью S_Φ для модели PLSA. Как видно из этих графиков, наибольшая разреженность достигается в модели ARTM1.

Анализ результатов построения тематических моделей показывает:

- для построения тематических моделей на данной выборке достаточно 20-30 итераций;
- для разбиения выборки на подвыборки лучше всего подходит модель ARTM, т.к. данная модель обладает наибольшей разреженностью S_Θ матрицы Θ, что

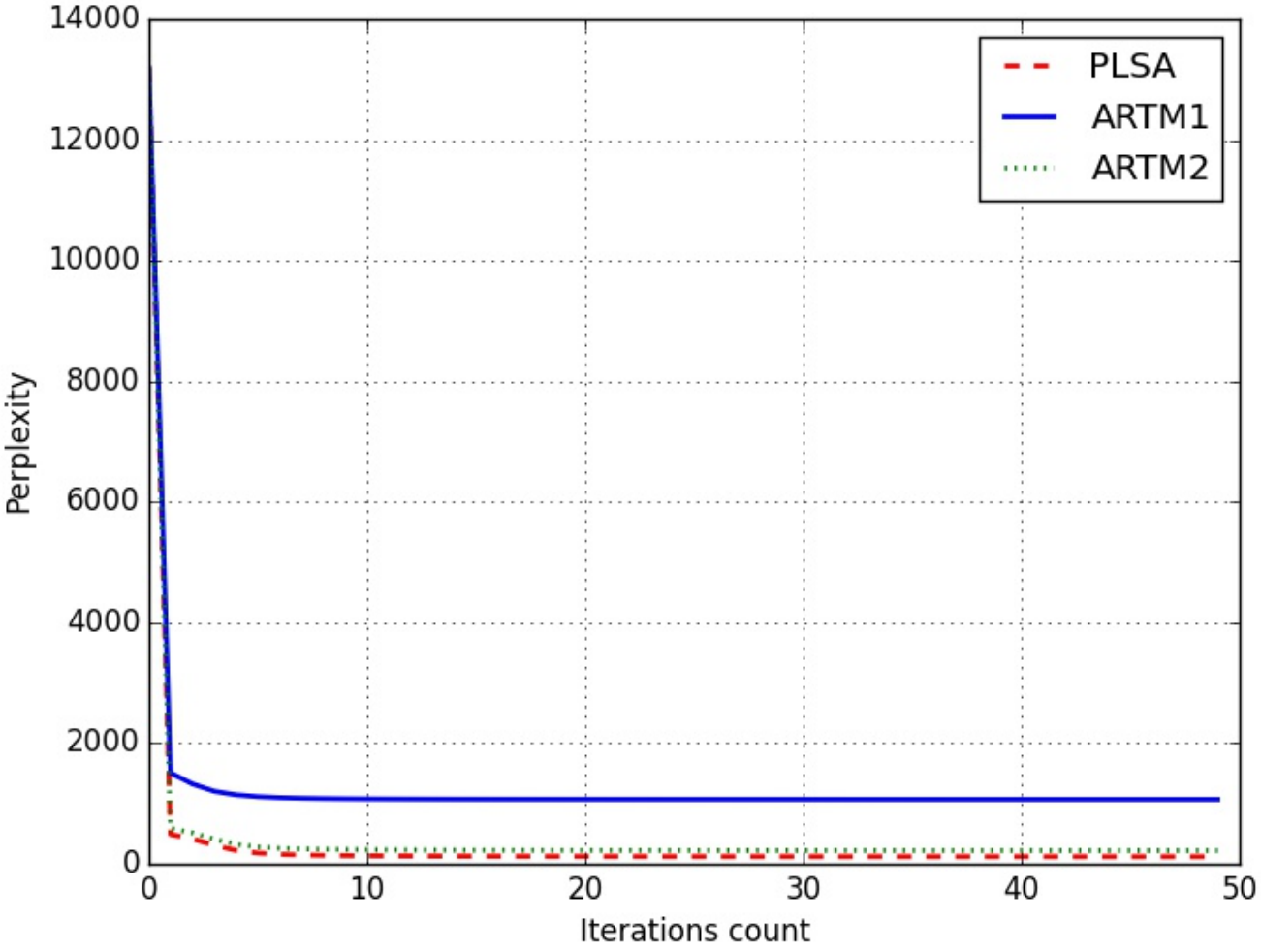

Рис. 3: Сравнение перплексии для моделей PLSA, ARTM1 и ARTM2

позволяет однозначно отнести закупки в подвыборки, кроме того, ухудшение перплексии по сравнению с моделью PLSA на данной модели сравнительно невелико;

- модель ARTM1 также подходит для разбиения выборки на подвыборки т.к. данная модель обладает большой разреженностью S_Θ матрицы Θ, однако, ухудшение перплексии по сравнению с моделью PLSA на данной модели достаточно велико, т.е. данная модель недостаточно хорошо предсказывает появление терминов $w \in W$ в текстах закупок $x \in \mathcal{D}$ и, как следствие, недостаточно хорошо объясняет данные.

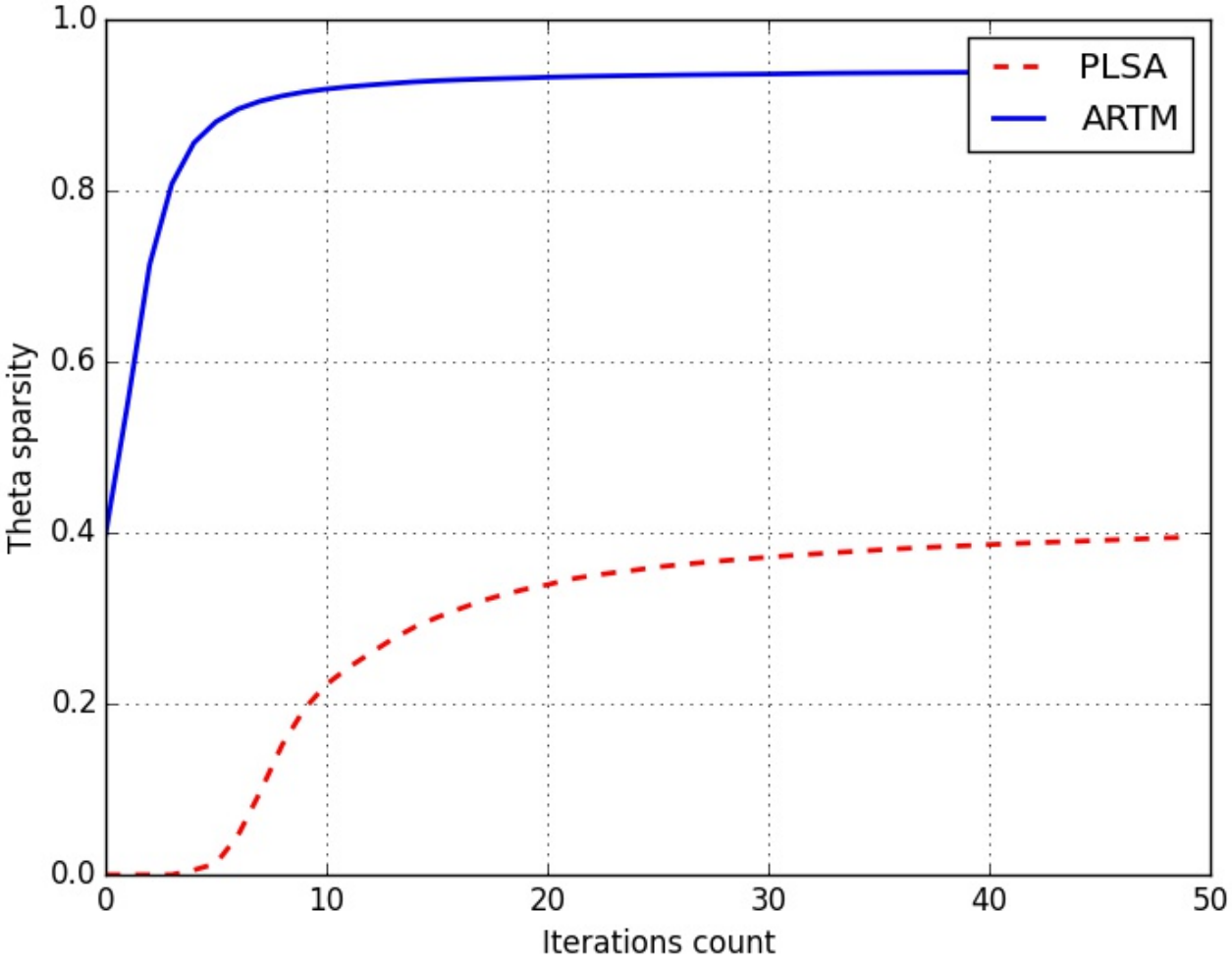

Рис. 4: Сравнение разреженности матрицы Θ для моделей PLSA и ARTM

- модели PLSA и ARTM2 не подходят для формирования подвыборок, т.к большой разреженности S_Θ матрицы Θ данных моделей недостаточно для однозначного отнесения закупки к определенной подвыборке.

Во второй части вычислительного эксперимента было проведено сравнение нескольких алгоритмов кластеризации с предложенной модификацией Hierarchical DBScan. К полученным в результате применения тематического моделирования с аддитивной регуляризацией подвыборкам были применены следующие алгоритмы кластеризации:

- Hierarchical DBScan;
- DBScan;

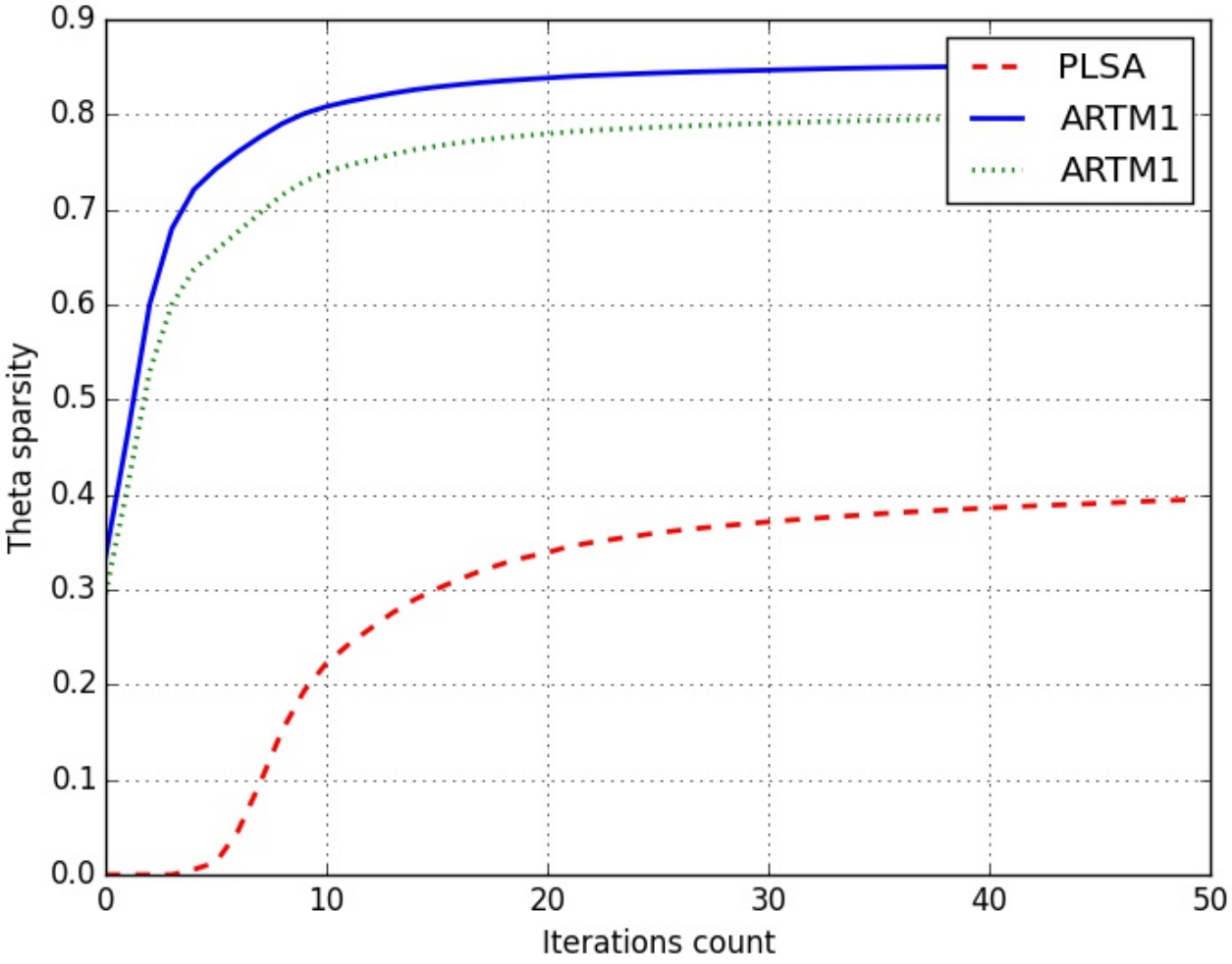

Рис. 5: Сравнение разреженности матрицы Θ для моделей PLSA, ARTM1 и ARTM2

- K-Means;
- Spectral Clustering;
- Affinity Propagation;
- Agglomerative Clustering;
- EM-алгоритм для декомпозиции смеси распределений;
- Topic Modeling.

Описание алгоритма Hierarchical DBScan и способо его применения описаны выше (см. раздел «Алгоритм решения поставленной задачи»). Для сравнения также были проведены эксперименты с оригинальной версией алгоритма DBScan; при этом веса

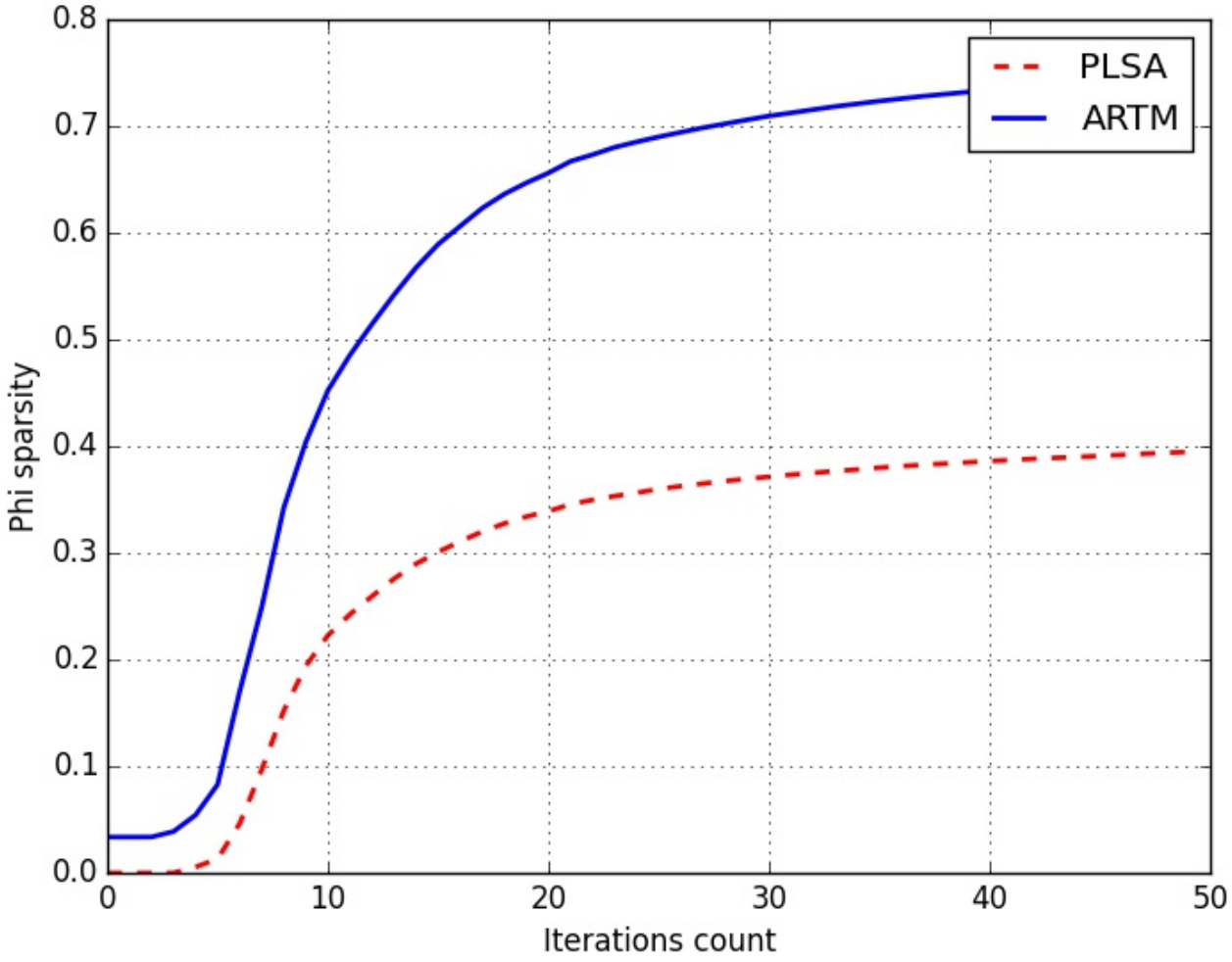

Рис. 6: Сравнение разреженности матрицы Φ для моделей PLSA и ARTM

в $\omega_1, \omega_2, \omega_3$ в функции расстояния подбирались по сетке значений, как и для алгоритма Affinity Propagation. В алгоритме K-Means, помимо значений весов, подбиралось также значение количества кластеров, как и в алгоритмах Agglomerative Clustering и Spectral Clustering. С помощью EM-алгоритма для декомпозиции смеси распределений задача кластеризации решалась как задача декомпозиции смеси распределений (не экспертно-интерпретируемой). При этом веса в смеси полагались не строго равными 0 или 1 (в нашей постановке задачи веса распределнений $w_C = [\mathbf{x} \in C]$). В Topic Modeling в качестве кластеров рассматривались полученные в первой части эксперимента подвыборки и проверялась гипотеза о том, что после применения тематического моделирования дополнительная кластеризация не нужна.

Результаты сравнения алгоритмов кластеризации по обоим критериям, представлен-

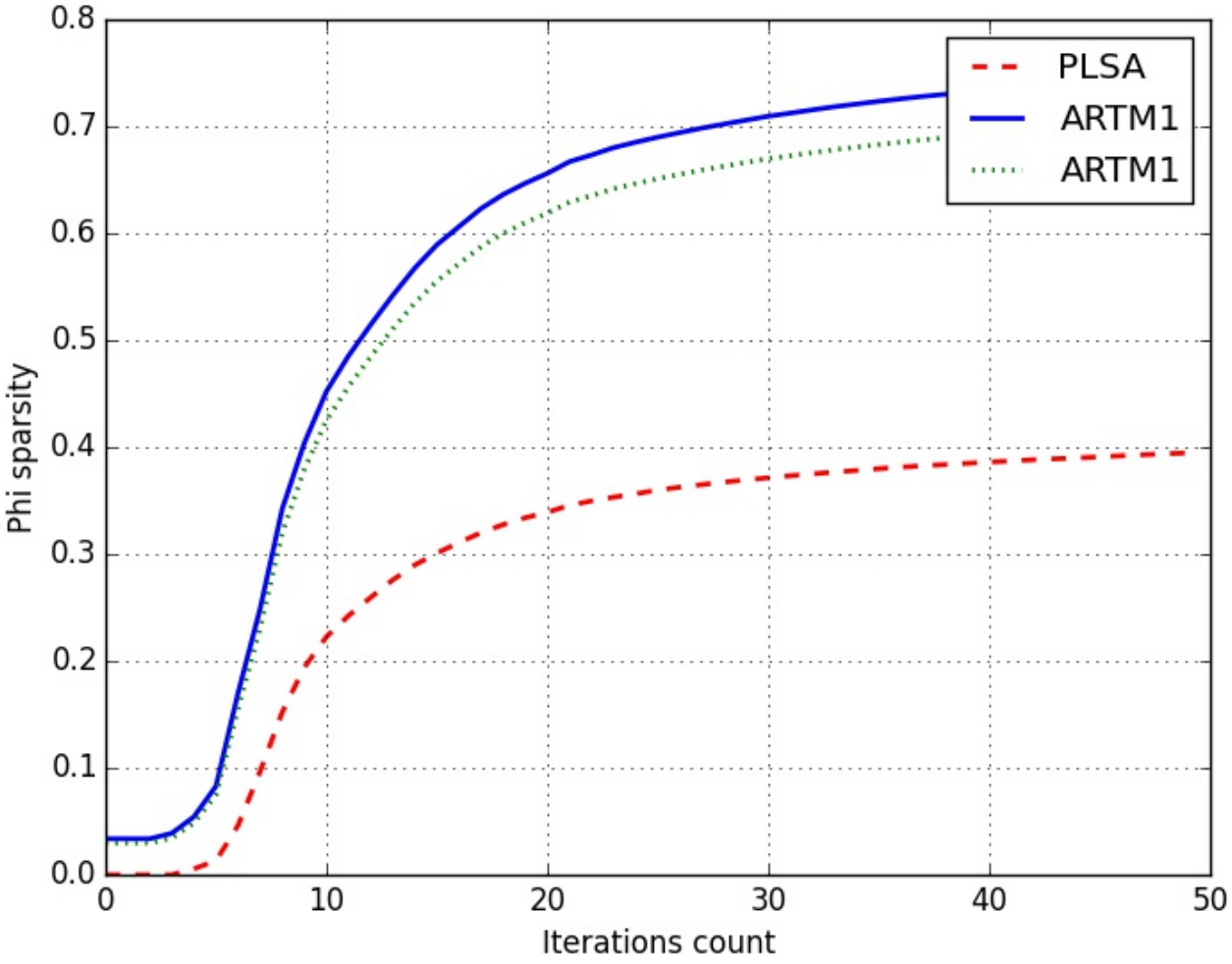

Рис. 7: Сравнение разреженности матрицы Φ для моделей PLSA, ARTM1 и ARTM2

ным в разделе «Постановка задачи», показаны в Табл. 2.

Анализ результатов построения кластеризации показывает:

- гипотеза о том, что после применения тематического моделирования дополнительная кластеризация не нужна была опровергнута;
- алгоритм DBScan без иерархической модификации показывает недостаточно хорошие для использования результаты;
- EM-алгоритм наилучшим образом оптимизирует экспертно-интерпертируемое правдоподобие, но в алгоитме учитывается как признак только цена, из-за чего экспертное агрегированное качество кластеров сравнительно невелико;
- результат работы алгоритма Hierarchical DBScan является парето-оптимальным

Таблица 2: Сравнение алгоритмов кластеризации

Алгоритм	Экспертно-интерпретируемое отрицательное лог-правдоподобие	Экспертное агрегированное качество кластеров
Hierarchical DBScan	1,14	0,79
DBScan	2,11	0,41
K-Means	1,52	0,75
Spectral Clustering	1,27	0,82
Affinity Propagation	1,87	0,61
Agglomerative Clustering	1,09	0,49
EM-алгоритм	0,93	0,21
Topic Modeling	5,27	0

и оптимальным в упрощенной постановке задачи при значениях параметра θ^{expert} меньших, чем 1,27;

- результаты работы алгоритмов Spectral Clustering, Agglomerative Clustering и EM-алгоритма также парето-оптимальны (см. Рис. 8);
- результаты работы алгоритма K-Means близки к результатам работы парето-оптимальных алгоритмов;
- алгоритм Hierarchical DBScan не требует точного подбора значений параметров, в отличие от остальных алгоритмов.

На рис. 9 и 10 показаны графики изменения экспертно-интерпертируемого отрицательного лог-правдоподобия и экспертного агрегированного качества кластеров в зависимости от максимально допустимого уровня иерархии. Как видно из графиков, 10 уровней иерархии достаточно.

Также в ходе вычислительного эксперимента была принята гипотеза о гамма-распределении цен внутри кластера.

5 Заключение

В данной работе:

- был представлен алгоритм кластеризации базы государственных закупок;

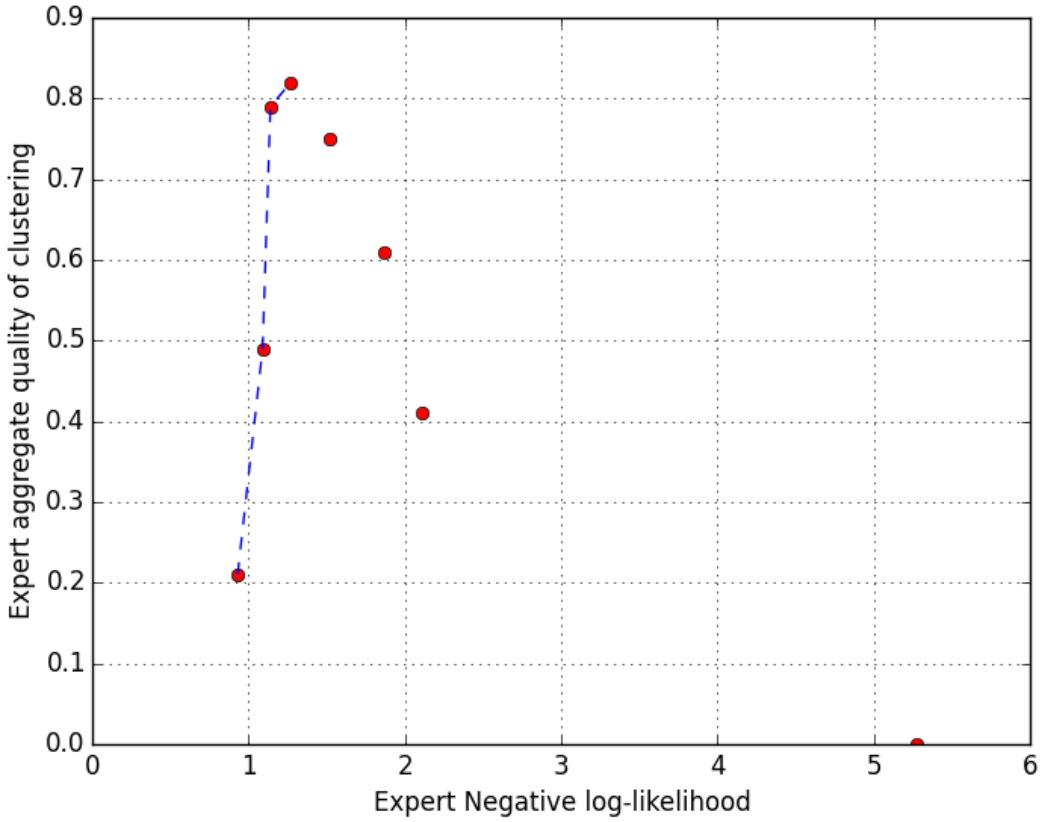

Рис. 8: Результаты сравенения алгоритмов кластеризации

- был представлен способ оценки рыночной цены товара;
- был представлен способ оценки типичности государственных закупок;
- был представлен способ оценки превышения рыночной цены;
- был представлен способ разбиения базы государственных закупок на непересекающиеся однородные подвыборки;
- было проведено сравнение представленного алгоритма кластеризации с несколькими существующими;
- была принята гипотеза о гамма-распределении цен внутри кластера.

Результаты данной работы использованы при разработке системы анализа государственных закупок «Антирутина-44».

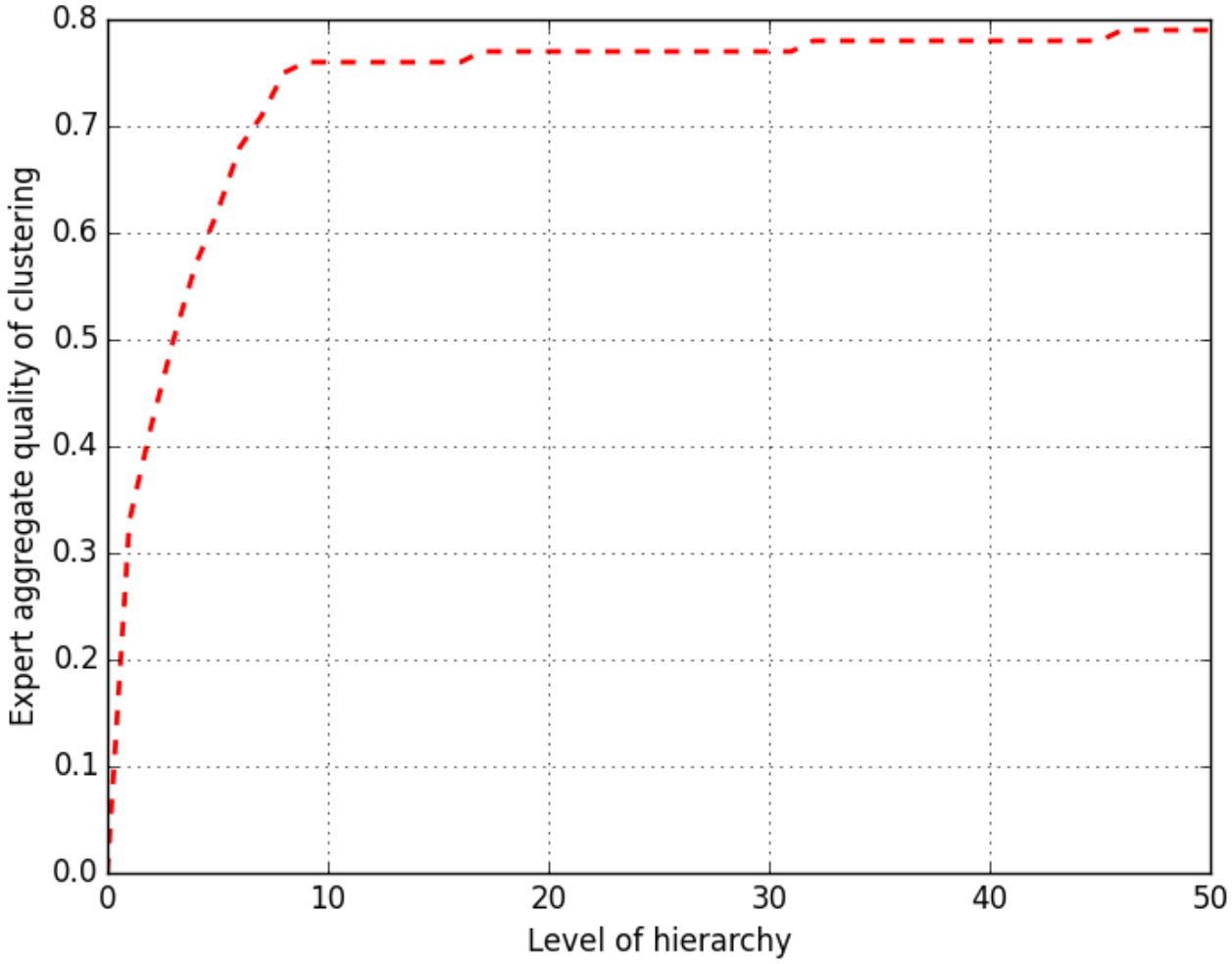

Рис. 9: Зависимость экспертного агрегированного качества кластеров от максимально допустимого уровня иерархии

Список литературы

[1] *Hong L., Davison B. D.* Empirical study of topic modeling in twitter // Proceedings of the first workshop on social media analytics. – ACM, 2010. – pp. 80-88.

[2] *Alvarez-Melis D., Saveski M.* Topic Modeling in Twitter: Aggregating Tweets by Conversations // Tenth International AAAI Conference on Web and Social Media. – 2016.

[3] *Rosa K. D. et al.* Topical clustering of tweets //Proceedings of the ACM SIGIR: SWSM. – 2011.

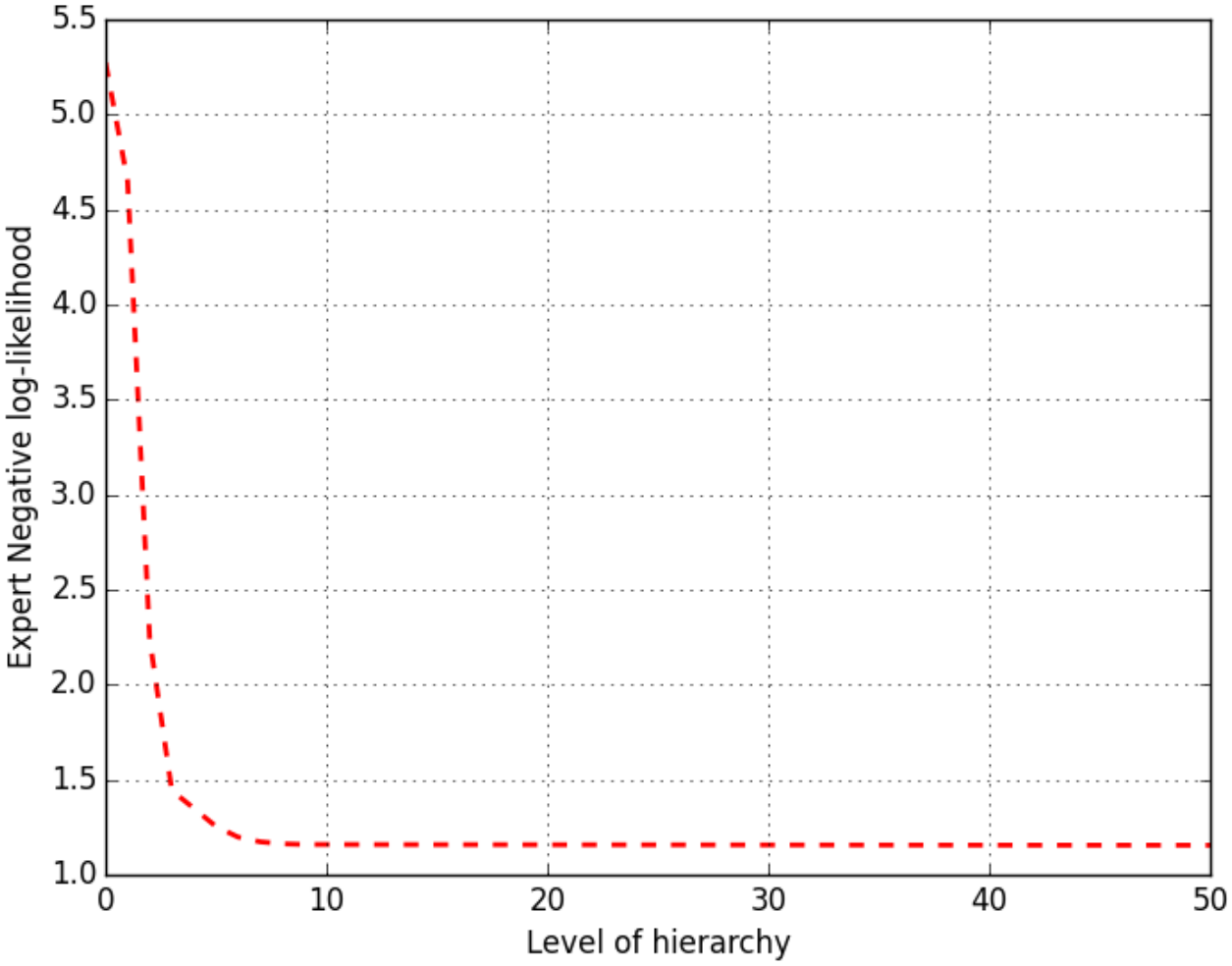

Рис. 10: Зависимость экспертно-интерпертируемого отрицательного лог-правдоподобия от максимально допустимого уровня иерархии

[4] *Tripathy R. M. et al.* Theme Based Clustering of Tweets // Proceedings of the 1st IKDD Conference on Data -Sciences. - ACM, 2014. – pp. 1-5.

[5] *Karandikar A.* Clustering short status messages: A topic model based approach : Diss. – University of Maryland, 2010.

[6] *Воронцов К. В.* Вероятностное тематическое моделирование // 2014.

[7] *Воронцов К. В.* Аддитивная регуляризация тематических моделей коллекций текстовых документов // Доклады РАН, 2014. — Т. 455., №3. 268–271

[8] *Воронцов К. В., Фрей А. И., Апишев М. А., Ромов П. А., Янина А. О., Суворова М. А.* BigARTM: библиотека с открытым кодом для тематического модели-

рования больших текстовых коллекций // Аналитика и управление данными в областях с интенсивным использованием данных. XVII Международная конференция DAMDID/RCDL'2015, Обнинск, 13-16 октября 2015.

[9] *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 // Topic Models: Post-Processing and Applications, CIKM 2015 Workshop, October 19, 2015, Melbourne, Australia.

[10] *Vorontsov K. V., Potapenko A. A., Plavin A. V.* Additive Regularization of Topic Models for Topic Selection and Sparse Factorization // The Third International Symposium On Learning And Data Sciences (SLDS 2015). April 20-22, 2015. Royal Holloway, University of London, UK. Springer International Publishing Switzerland 2015, A. Gammerman et al. (Eds.): SLDS 2015, LNAI 9047, pp. 193–202, 2015.

[11] BigARTM's documentation // http://docs.bigartm.org/en/stable/index.html

[12] *Ester M. et al.* A density-based algorithm for discovering clusters in large spatial databases with noise //Kdd. – 1996. – Vol. 96., №. 34. – pp. 226-231.

[13] *Campello R. J. G. B., Moulavi D., Sander J.* Density-based clustering based on hierarchical density estimates //Advances in Knowledge Discovery and Data Mining. – Springer Berlin Heidelberg, 2013. – pp. 160-172.

[14] *Lloyd S.* Last square quantization in pcm's //Bell Telephone Laboratories Paper. – 1957.

[15] *MacQueen J. et al.* Some methods for classification and analysis of multivariate observations //Proceedings of the fifth Berkeley symposium on mathematical statistics and probability. – 1967. – Vol. 1., №. 14., pp. 281-297.

[16] *Frey B. J., Dueck D.* Clustering by passing messages between data points // science. – 2007. – Vol. 315., №. 5814., pp. 972-976.

[17] *Shi J., Malik J.* Normalized cuts and image segmentation // Pattern Analysis and Machine Intelligence, IEEE Transactions on. – 2000. – Vol. 22, №. 8, pp. 888-905.

[18] *Von Luxburg U.* A tutorial on spectral clustering // Statistics and computing. – 2007. – Vol. 17., №. 4., pp. 395-416.

[19] *Sibson R.* SLINK: an optimally efficient algorithm for the single-link cluster method // The Computer Journal. – 1973. – Vol. 16., №. 1., pp. 30-34.

[20] *Defays D.* An efficient algorithm for a complete link method //The Computer Journal. – 1977. – Vol. 20., №. 4., pp. 364-366.

[21] *Maimon O., Rokach L.* Data mining and knowledge discovery handbook. – New York : Springer, 2005. – Vol. 2.

[22] NLTK Documentation // http://www.nltk.org/

[23] Pymophy2 Documentation // https://pymorphy2.readthedocs.io

[24] Wikipedia // https://en.wikipedia.org/, https://upload.wikimedia.org/wikipedia/commons/thumb/a/
Illustration.svg/165px-DBSCAN-Illustration.svg.png