Аннотация

Работа посвящена разработке методов анализа потребительского поведения розничных клиентов банка. Ставится задача по нахождению моментов времени изменения потребительского поведения клиентов. Строится тематическая сегментация истории транзакций клиента на основе PLSA, LDA и ARTM моделей тематического моделирования. Получен метод векторного представления истории транзакций клиента. Разработан метод оценивания качества сегментации истории транзакций. Подтверждена гипотеза о незначительном изменении качества сегментации при переходе к тематическому представлению для искусственных историй транзакций.
Ключевые слова: тематическое моделирование; ARTM; сегментация; mcc-коды; Topic Tiling; транзакционные данные; потребительское поведение

Содержание

1 Введение

Анализ данных банковских транзакций розничных клиентов позволяет получить дополнительную информацию о клиенте. Такая информация полезна в аналитике активности клиентов и составлении их профиля потребительского поведения. Под профилем потребительского поведения клиента понимается оценка потребительского поведения клиента банка. Потребительское поведение характеризует, у каких продавцов клиент делает покупки. Профиль потребительского поведения клиента позволяет выделять категории клиентов со схожим поведением, прогнозировать крупные или характерные покупки отдельных клиентов, предоставлять релевантные предложения.

Код категории продавца (Merchant Category Code - mcc-код) - четырехзначный код, автоматически выдаваемый банком продавцу при регистрации терминала оплаты на основе информации об основном роде деятельности торговой точки. Таким образом, mcc-код транзакции не позволяет точно установить покупку, но в некоторых случаях несет важную информацию об отрасли товаров, в которой была сделана покупка. К примеру, mcc-код 5311 соответствует универмагам и слабо описывает покупку, а mcc-код 5641 соответствует точкам по продаже детской одежды, включая одежду для самых маленьких, что несет гораздо больше информации о характере покупки.

История банковских транзакций содержит информацию о mcc-кодах торговых точек, в которых эти транзакции были совершены. Поэтому по история транзакций можно построить профиль потребительского поведения клиента. Предполагается, что в потребительском поведении можно выделить темы, с помощью вектора которых можно охарактеризовать потребительское поведение любого клиента. Такие темы строятся с помощью тематического моделирования. В работе [1] была показана применимость подходов тематического моделирования к построению профиля розничных клиентов банка по транзакционным данным. В ней были построены PLSA [2], LDA [3] и ARTM [4] модели по транзакционным данным.

Со временем потребительское поведение клиента может изменяться. Изменения могут быть связанны с различными, запланированными или незапланированными, событиями в жизни клиента, например, планируемый отпуск или внеплановый ремонт автомобиля.

Определение подобных моментов изменения потребительского поведения позволяет улучшить качество профиля потребительского поведения клиента и своевременно делать релевантные персональные предло-

жения. Задачу поиска моментов изменения потребительского поведения клиентов можно переформулировать как задачу о нахождении границ однородного потребления, или участков с устойчивым во времени набором mcc-кодов.

Решение этой задачи можно разбить на два этапа.

Этап первый: Выделение множества тем потребительского поведения клиентов с помощью тематической модели построенной на транзакционных данных розничных клиентов. А затем представление транзакционных данных каждого клиента последовательностью тематических векторов.

Этап второй: Сегментирование последовательности векторов. В работе [5] решается схожая задача сегментации диалогов контактного центра. Ключевое отличие от нашей задачи - на втором этапе требуется выделение монотематических сегментов. В данной работе это ограничение отсутствует. В работе [6] для решения схожей задачи разбития текста на сегменты с помощью тематических векторов предлагается использовать алгоритм Topic Tiling.

Для разрешения второго этапа используется аналог алгоритма Topic Tiling [6], учитывающий специфику транзакционных данных.

Сложность работы заключается в отсутствии размеченных данных и отсутствии четких критериев изменения потребительских привычек.

Работа посвящена разработке методов анализа потребительского поведения розничных клиентов банка. В рамках работы проводится два эксперимента. Первый определяет качество проведения сегментов аналогом алгоритма Topic Tiling на искусственных профилях (границы сегментов в которых мы знаем). Второй эксперимент показывает, насколько результат тематической сегментации реальных профилей клиентов похож на обычную сегментацию. В обоих экспериментах алгоритм аналога Topic Tiling фиксирован и отличается только входными данными, полученными после проведения первого этапа.

В ходе проведения экспериментов решаются задачи получения векторных представлений истории транзакций клиента для проведения сегментации, разработки метода оценивания качества сегментации истории транзакций клиента и проверки гипотезы о незначительном изменении качества сегментации при переходе к тематическому представлению истории клиента.

2 Постановка Задачи

2.1 Тематическая модель: основные понятия и определения

В качестве векторных представлений транзакций пользователей могут выступать тематические векторы — векторы распределений тем в транзакции, полученные путем построения тематической модели по коллекции профилей пользователя.

Обозначим D - множество профилей пользователей (коллекция). Пусть W - множество mcc-кодов (словарь). Будем называть профилем пользователя $d \in D$ совокупность дополнительной информации о пользователе, такой как пол, возраст, образование, семейное положение и историю (последовательность) его транзакций $w_{d,1}, \ldots, w_{d,N_d}$, где N_d - число транзакций пользователя за выбранный период.

Предполагаем, что для определения характера потребления пользователя порядок транзакций в его истории не важен, а важен набор mcc-кодов и суммарная трата на каждый код (гипотеза о мешке слов). Обозначим n_{wd} - сумму транзакций клиента d по mcc-коду w. Тогда, можем заключить информацию об истории всех пользователей в матрицу частот mcc-кодов в истории пользователей F размера $|W| \times |D|$ состоящую из элементов n_{wd}.

Можем записать следующие частотные оценки вероятностей, получающиеся непосредственно из данных коллекции:

$$\hat{p}(d,w) = \frac{n_{dw}}{n}, \quad \hat{p}(d) = \frac{n_d}{n}, \quad \hat{p}(w) = \frac{n_w}{n}, \quad \hat{p}(w|d) = \frac{n_{dw}}{n_d}; \tag{2.1}$$

$$n_d = \sum_w n_{dw} - \text{общая сумма трат пользователя } d;$$

$$n_w = \sum_d n_{dw} - \text{сумма покупок всех пользователей по mcc-коду} w;$$

$$n = \sum_d \sum_w n_{dw} - \text{сумма трат всех пользователей.}$$

Полагаем, что наличие любого mcc-кода $w \in W$ в истории транзакций пользователя $d \in D$ связано с некой скрытой переменной $t \in T$. Где T - множество возможных составляющих характера поведения пользователя (темы).Тогда можем представить коллекцию D как выборку n_{tdw}

из $p(d, w, t)$ на множестве $D \times W \times T$. Где n_{tdw} - связанная с темой t сумма транзакций клиента d по mcc-коду w.

В качестве гипотезы условной независимости возьмем предположение о том, что вероятность наличия mcc-кода в истории пользователя d с характером потребления t зависит от характера t, но не зависит от пользователя d, то есть может быть описана общим для всех пользователей распределением $p(w|t)$:

$$p(w|d, t) = p(w|t). \tag{2.2}$$

Тогда с помощью формулы полной вероятности и гипотезы условной независимости можем описать распределение mcc-кодов у пользователя $p(w|d)$ вероятностной смесью распределений mcc-кодов в характерах потребления $\varphi_{wt} = p(w|t)$ с весами $\theta_{td} = p(t|d)$:

$$p(w|d) = \sum_{t \in T} p(w|t, d)p(t|d) = \sum_{t \in T} p(w|t)p(t|d) = \sum_{t \in T} \varphi_{wt}\theta_{td}. \tag{2.3}$$

Вероятностная модель (2.3) описывает процесс порождения истории транзакций по известным распределениям $p(w|t)$ и $p(t|d)$.

Задача тематического моделирования — это обратная задача: по заданному множеству профилей D требуется найти параметры φ_{wt} и θ_{td}, при которых тематическая модель (2.3) хорошо приближает частотные оценки условных вероятностей $\hat{p}(w|d) = \frac{n_{wd}}{n_d}$.

Выпишем частотные оценки вероятностей связанных с характером потребления t:

$$\hat{p}(t) = \frac{n_t}{n}, \quad \hat{\varphi}_{wt} = \hat{p}(w|t) = \frac{n_{wt}}{n_t}, \quad \hat{\theta}_{td} = \hat{p}(t|d) = \frac{n_{td}}{n_d}, \quad \hat{p}(t|d, w) = \frac{n_{tdw}}{n_{dw}}; \tag{2.4}$$

$n_{td} = \sum_w n_{tdw}$ – связанная с темой t сумма транзакций клиента d;

$n_{wt} = \sum_d n_{tdw}$ – связанная с темой t сумма транзакций по mcc-коду w;

$n_t = \sum_d \sum_w n_{tdw}$ – сумма трат всех пользователей связанная с темой t.

Простейшей вероятностной тематической моделью является модель PLSA [2]. В PLSA для построения модели (2.3) максимизируется логарифм правдоподобия (2.5) при ограничениях нормировки и неотрица-

тельности (2.6) :

$$L(\Phi,\Theta)=\sum_{d\in D}\sum_{w\in d}n_{dw}\ln\sum_{t\in T}\varphi_{wt}\theta_{td}\ \to\ \max_{\Phi,\Theta};\tag{2.5}$$

$$\sum_{w\in W}\varphi_{wt}=1;\quad \varphi_{wt}\geq 0;\qquad \sum_{t\in T}\theta_{td}=1;\quad \theta_{td}\geq 0.\tag{2.6}$$

Равенство (2.3) представляется в матричном виде. В левой части равенства находится известная матрица частот MCC-кодов у клиентов $F=\big(\hat{p}(w|d)\big)_{W\times D}$. Правая часть представляет собой произведение двух неизвестных матриц — матрицы $\Phi=(\varphi_{wt})_{W\times T}$ и матрицы $\Theta=(\theta_{td})_{T\times D}$. Считаем, что $|T|$ много меньше $|D|$ и $|W|$, поэтому задача тематического моделирования сводится к поиску приближённого матричного разложения $F\approx\Phi\Theta$, ранг которого не превышает $|T|$.

Задача стохастического матричного разложения является некорректно поставленной, так как множество её решений в общем случае бесконечно. Если $\Phi\Theta$ — решение, то $(\Phi S)(S^{-1}\Theta)$ также является решением для всех невырожденных матриц S, при условии, что матрицы ΦS и $S^{-1}\Theta$ — стохастические.

Аддитивная регуляризация тематических моделей (ARTM) [4] основана на максимизации линейной комбинации логарифма правдоподобия и нескольких регуляризаторов $R_i(\Phi,\Theta),\ i=1,\ldots,k$:

$$\sum_{d\in D}\sum_{w\in d}n_{dw}\ln\sum_{t\in T}\varphi_{wt}\theta_{td}+R(\Phi,\Theta)\ \to\ \max_{\Phi,\Theta};\qquad R(\Phi,\Theta)=\sum_{i=1}^{k}\tau_i R_i(\Phi,\Theta);\tag{2.7}$$

при ограничениях (2.6), где τ_i — неотрицательные коэффициенты регуляризации.

Регуляризатор сглаживания и разреживания матриц Φ,Θ:

$$R(\Phi,\Theta)=\sum_{t\in T}\sum_{w\in W}\beta_{wt}\ln\varphi_{wt}+\sum_{d\in D}\sum_{t\in T}\alpha_{td}\ln\theta_{td}\tag{2.8}$$

Положительные значения β_{wt} и α_{td} соответствуют сглаживанию распределений φ_{wt} и θ_{td}, а отрицательные - разреживанию. Предполагается, что у клиента проявляется малое количество тем, что соответствует разряженности матрицы Θ.

Описанная в [3] модель LDA (Latent Dirihlet Alloation) опирается на предположение, о том, что вектора θ_d и φ_t случайны и получены из распределения Дирихле.

Согласно работе [7], модели LDA получается применением регуляризатора, равному с точностью до константы логарифму априорного распределения Дирихле:

$$R(\Phi,\Theta) = \ln \prod_{t\in T} Dir(\varphi_t;\beta) \prod_{d\in D} Dir(\theta_d;\alpha) + const = \sum_{t\in T}\sum_{w\in W} (\beta_w - 1)\ln \varphi_{wt} + \sum_{d\in D}\sum_{t\in T} (\alpha_t - 1)\ln \theta_{td}$$

2.2 Тематика mcc-кода и транзакции

Зная $\hat{\varphi}_{wt}$ и $\hat{\theta}_{td}$ с помощью (2.1) и (2.4) можем оценить распределение $p(t|w)$:

$$\hat{p}(t|w) = \frac{p(w|t)p(t)}{p(w)} = \frac{\hat{\varphi}_{wt}\sum_d \hat{\theta}_{td} n_d}{n_w} \tag{2.9}$$

Распределение $p(t|w)$ называют *тематикой mcc-кода* w.

С помощью (2.1) , (2.4) и гипотезы условной независимости можем оценить распределение $p(t|d,w)$

$$\hat{p}(t|d,w) = \frac{\hat{p}(t,w|d)}{\hat{p}(w|d)} = \frac{\hat{p}(w|t,d)\hat{p}(t|d)}{\hat{p}(w|d)} = \frac{\hat{p}(w|t)\hat{p}(t|d)}{\hat{p}(w|d)} = \frac{\hat{\varphi}_{wt}\hat{\theta}_{td}}{\sum_s \hat{\varphi}_{ws}\hat{\theta}_{sd}} \tag{2.10}$$

Распределение $p(t|d,w)$ будем называть *тематикой транзакции* w *клиента* d.

Таким образом, определены простейшая тематическая модель PLSA, и модель с регуляризацией ARTM, построенные по транзакционным данным. С помощью задачи тематического моделирования, определены понятия тематики mcc-кода $p(t|w)$ и тематики транзакции $p(t|d,w)$ для получения тематического представления истории транзакций клиента.

2.3 Тематическое представление истории транзакций

Для тематического векторного представления истории транзакций клиента используются различные подходы к представлению отдельных транзакций.

Первый подход: представление транзакции вектором с единицей на месте идентификатора mcc-кода (*one-hot представление*). По истории транзакций представленной в таком виде можно однозначно восстановить исходный порядок mcc-кодов в истории транзакций клиента. Значит данный подход не теряет информацию о последовательности действий клиента. Однако он не использует информацию о суммах транзакций. Данный подход используется для построения *простой сегментации* истории транзакций клиента (без применения тематического моделирования).

Второй подход: представление транзакций тематикой mcc-кода транзакции (2.9)

Третий подход: представление транзакций тематикой транзакции (2.10)

По истории транзакций представленной с помощью второго и третьего подхода нельзя однозначно восстановить исходный порядок mcc-кодов в истории транзакций клиента. Значит, часть информации о последовательности действий клиента может быть утеряна. Третий подход использует информацию о суммах транзакций.

Выдвигается гипотеза незначительного изменения качества сегментации при переходе от пиостой сегментации (первый подход) к сегментации с использованием тематического векторного представления (второй и третий подход).

2.4 Сегментация транзакционных данных

Назовем *гистограммным вектором* вектор размерности $|W|$, созданный на основе гистограммы количества mcc-кодов в наборе. Значение компоненты вектора, отвечающей соответствующему mcc-коду, равно количеству вхождений этого кода в набор.

Будем говорить, что два набора mcc-кодов *похожи*, если косинусная мера между их гистограммными векторами превышает некий порог. Предполагаем, что история транзакций каждого клиента обладает своим порогом похожести.

Потребительское поведение *однородно на отрезке времени* $[s_{d,i}, s_{d,i+1}]$, если на любых двух временных отрезках одинаковой длинны, содержащихся в $[s_{d,i}, s_{d,i+1}]$, набор mcc-кодов похож.

Сегментом однородного потребительского поведения клиента назовем временной отрезок, при любом расширении которого свойство однородности потребительского поведения теряется.

Задача сегментации истории транзакций клиента

Таким образом, задача сегментации истории транзакций клиента состоит

в определении m границ сегментов однородного потребительского поведения $s_{d,1}, \ldots, s_{d,m-1}$ по последовательности транзакций клиента $d \in D$ в тематическом векторном представлении $w_{d,1}, \ldots, w_{d,n_d}$.

2.5 Аналог Topic Tiling

Задачу о сегментации истории транзакций клиента предлагается решать с помощью аналога алгоритма Topic Tiling [6]. Описание алгоритма представлено ниже:

- На вход подается последовательность транзакций клиента (профиль), в тематическом векторном представлении виде (рис.(1) (а)).
- Проходим по профилю двумя скользящими окнами $(i - h_1, h_1]$ и $(h_1, i + h_1]$, для каждого i вычисляя косинусную меру между средним левого и правого окна (рис.(1) (б) cosine_similarity).
- Строится график косинусной меры от i и сглаживается окном h_2 (рис.(1) (б) smoothed_cosine_similarity).
- Отмечаются все локальные минимумы на графике косинусной меры - это потенциальные места для проведения границ сегмента.
- Для всех локальных минимумов считается уверенность проведения сегмента в i:

  $$depth\_score(i) = \frac{1}{2}(hl(i) - c_i + hr(i) - c_i),$$

  где c_i - значение косинусной меры в i, $hl(i)$ - ближайший к i локальный максимум слева, а $hr(i)$ - справа (рис.(1) (б) depth_score).
- Если количество сегментов задано как n, то выберается $n - 1$ максимальных по значению $depth\_score(i)$ мест и в них проводится границы сегментов (рис.(1) (б) segment border).
- Если количество сегментов не задано, то определяется $threshold$и проводятся границы везде, где $depth\_score(i) > threshold$ (рис.(1) (б) segment border).

Отличие данного алгоритма от [6] в двух местах: во-первых, добавлено сглаживание графика косинусной меры перед нахождением локальных экстремумов, и, во-вторых, изменен порог принятия решения о выставлении границы сегмента.

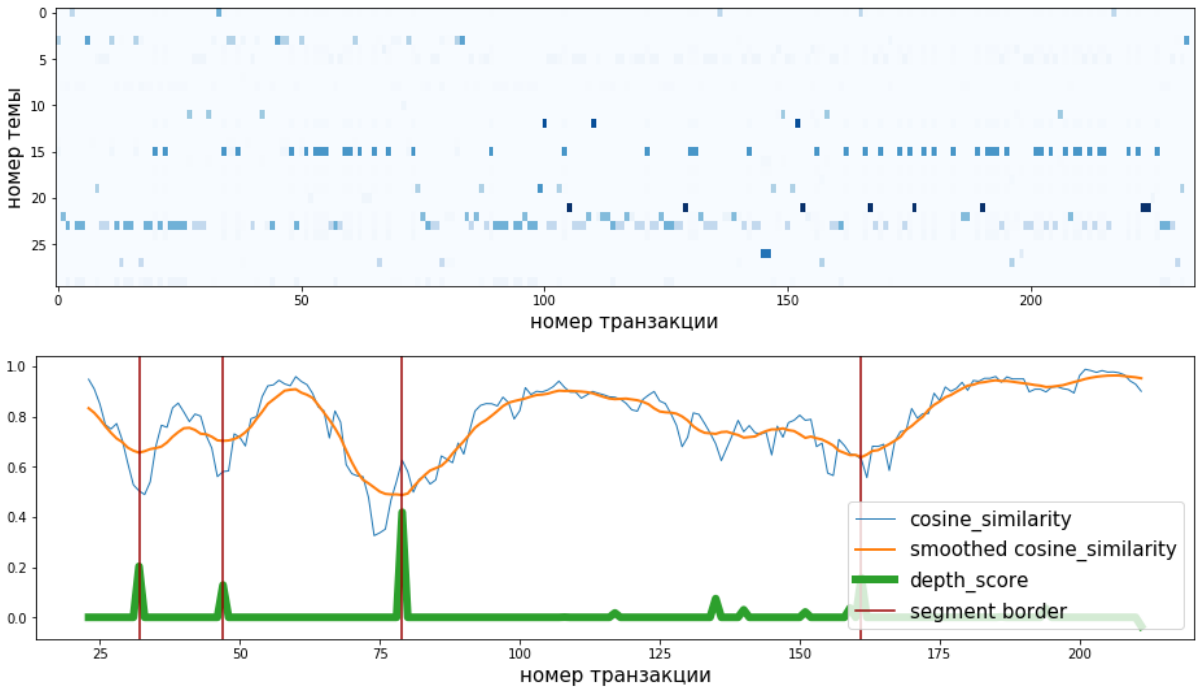

Рис. 1: а) Векторное представление одного профиля (сверху).
б) Стадии работы Аналога Topic Tiling (снизу)

2.6 Оценка качества сегментации

Для оценки качества проведения границ сегментов используется метрика ошибок сегментации текстов естественного языка P_k - метрика, введенная в работе [8]. Так же используется метрика ошибок Window Diff, предложенная, как альтернатива P_k - метрике, в работе [9]. Обе метрики вычисляются для каждого профиля $d \in D$ и основаны на сравнении сегментации, качество которой необходимо оценить(*предсказанная сегментация*) и некоторой сегментацией-образцом(*истинная сегментация*).

2.6.1 P_k метрика

Для подсчета P_k метрики определим *невязку* между истинной и предсказанной сегментацией.
Для каждого профиля рассматриваются пары транзакций:

$$(w_{d,1}, w_{d,k+1}), \ldots, (w_{d,N_d-k}, w_{d,N_d})$$

Для каждой пары записывается 0, если они находятся в одном сегменте и 1, если в разных.
Невязка для P_k метрики, $b_{s_d}(i)$, - доля несовпадений между значениями

предсказанной и истинной сегментацией:

$$b_{s_d}(i) = [w_{d,i} \in s_{d,q}][w_{d,i+k} \in s_{d,q}],$$

где $s_{d,q}$ - некий сегмент сегментации s_d профиля d.
P_k метрика для профиля d, $P_k(d)$, выражается через невязку $b_{s_d}(i)$ следующим образом:

$$P_k(d) = \frac{1}{n_d - k} \sum_{i=1}^{N_d - k} [b_{s_{d,true}}(i) \neq b_{s_d}(i)]$$

2.6.2 Window Diff метрика

В отличие от P_k - метрики, Window Diff сопоставляет каждой паре транзакций не 1 или 0, а количество границ сегментов, находящихся между транзакциями данной пары. *Невязка для Window Diff*, $b_{s_d}(i)$, - доля несовпадений между значениями предсказанной и истинной сегментацией:

$$\begin{cases} w_{d,i} \in s_{d,q} \\ w_{d,i+k} \in s_{d,t} \end{cases} \implies b_{s_d}(i) = t - q$$

где $s_{d,q}$ и $s_{d,t}$ - некие сегменты сегментации s_d профиля d с порядковыми номерами q и t соответственно.
Window Diff метрика для профиля d, WindowDiff(d), выражается через невязку $b_{s_d}(i)$ следующим образом:

$$\text{WindowDiff}(d) = \frac{1}{N_d - k} \sum_{i=1}^{N_d - k} [b_{s_{d,true}}(i) \neq b_{s_d}(i)]$$

2.7 Формальная постановка задачи

Дано: $< w_i, \tau_i, s_i >_{i=1}^{n_d}$ - история транзакций клиента d, где
w_i - mcc-код продавца
τ_i - дата и время транзакции
s_i - сумма транзакции в рублях
Найти:

1. моменты изменения потребительского поведения (границы сегментов)

2. тематическое векторное представление истории транзакций клиента

3. рамки применимости гипотезы о незначительном изменении качества сегментации при переходе к тематическому представлению истории транзакций клиента

Критерием качества сегментации является P_k и $WindowDiff$ метрики. Критерием качества векторного представления клиента является качество сегментации проведенной с помощью этого векторного представления.

3 Построение моделей, вычислительные эксперименты

3.1 Описание и обработка транзакционных данных

Дан набор пользователей D. В работе использованы данные о всех транзакциях для набора пользователей D за три года. Транзакции записаны в таблице, имеющей поля:

- mcc_code - mcc-код транзакции
- $amount\_ru$ - сумма транзакции в рублях
- $cardnumber$ - уникальный идентификатор карты пользователя
- $trans\_time$ - дата и время транзакции (с точностью до секунды),

где сумма транзакции указана со знаком, означающим списание или зачисление средств. В эксперименте использовались только данные о списаниях,то есть были взяты толькотранзакции с $amount\_ru < 0$. Кроме того, были удалены все транзакции с mcc-кодами финансовых операций и некоторые частые mcc-коды, такие как коды покупок в супермаркетах.

Для каждого пользователя была выделена и упорядочена во времени история его транзакций (профиль).

3.2 Описание моделей

Опишем модели для задания тематических векторных представлений историй транзакций, используемые в экспериментах.

Модели без применения тематического моделирования

- **one-hot** - one-hot представление.
- **random** - тематическое представление каждого mcc-кода задано вектором, сгенерированным из равномерного распределения.
- **lazy** - тематическое представление всех транзакций одинаково (модель не проводит границы сегментов).

Модели на основе тематической модели

Каждая из данных моделей использует тематики, полученные с помощью тематического моделирования. В экспериментах рассмотрены следующие модели:

- **PLSA**
- **LDA**
- **ARTM_SmoothTheta** - ARTM модель с регуляризатором сглаживающим матрицу Θ
- **ARTM** модель с модальностями и набором регуляризаторов (субьективно лучшая по качеству тем).

Модели различаются числом тем. По-умолчанию, все модели построены на 30 темах. Если в названии модели фигурирует число (например,

3.2.1 Поиск границ сегментов

Цель данного эксперимента — определить качество сегментации искусственных профилей аналогом алгоритма Topic Tiling для тематических векторных представлений транзакций различных моделей.

Полагаем, что похожесть(раздел 2.4) между частями историй транзакций разных клиентов меньше, чем похожесть между частями внутри истории транзакций одного клиента.

Создадим несколько искусственных профилей, склеив части реально существующих историй транзакций. Границами истинной сегментации будем считать границы склеивания частей сегментов различных профилей.

	P_k		WindowDiff	
Модель	mean	std	mean	std
one-hot	**0.069**	**0.057**	**0.069**	**0.057**
LDA	0.079	0.076	0.080	0.077
PLSA	0.084	0.081	0.084	0.082
PLSA_45	0.077	0.070	0.078	0.072
PLSA_80	0.081	0.067	0.081	0.068
ARTM	0.192	0.128	0.212	0.147
Random	0.118	0.109	0.128	0.118
ARTM_SmoothTheta	0.088	0.074	0.088	0.074

Таблица 1: Точность проведения сегментов по сравнению с истинными на 300 искусственных профилях

Сравнение моделей

В таблице (1) и на графике (2) представлена ошибка проведения сегментов по сравнению с истинными границами сегментов на искусственной

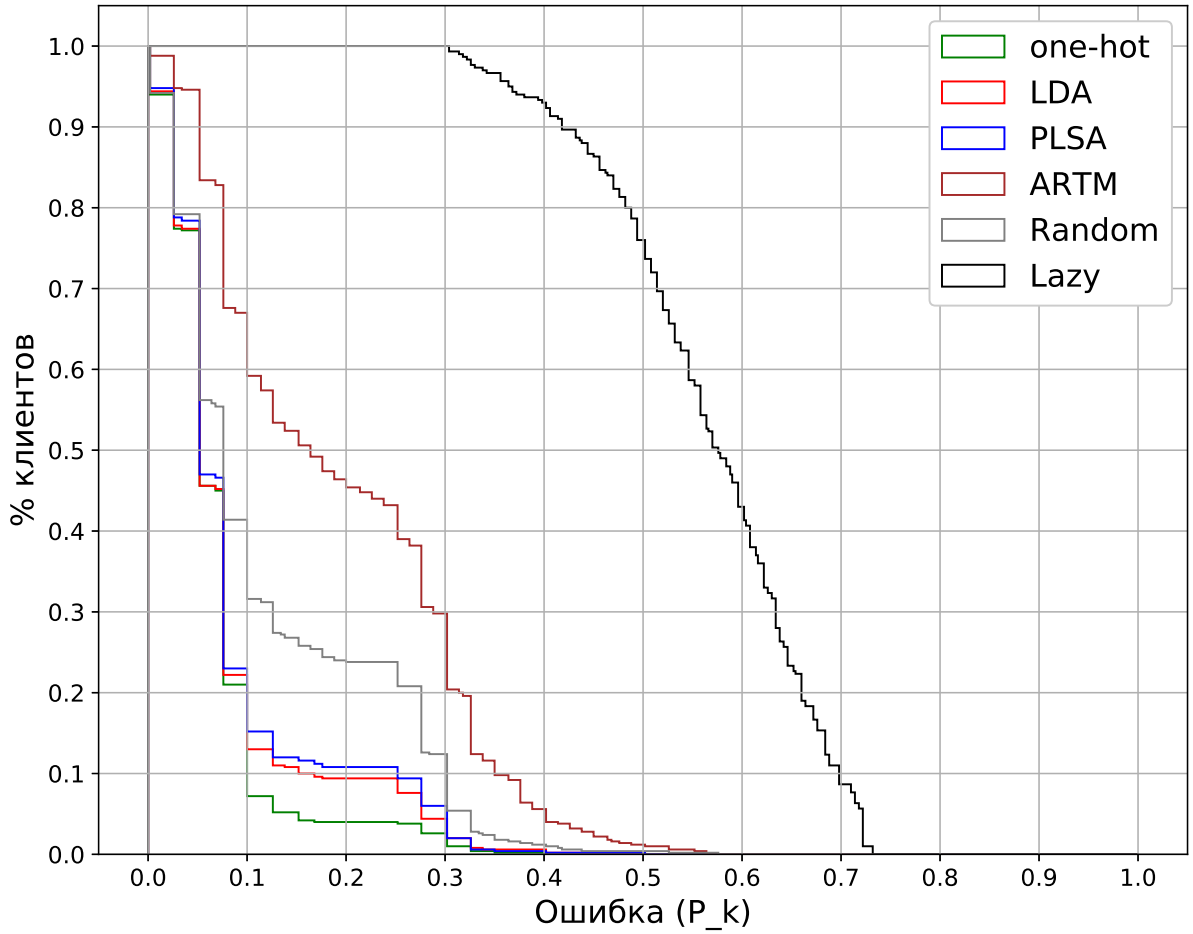

Рис. 2: Точность проведения сегментов по сравнению с истинными на 300 искусственных профилях

коллекции для различных моделей генерации тематических векторных представлений. Ошибка измерена с помощью P_k и $WindowDiff$ метрик для 300 искусственных профилей. В таблицу занесены mean (выборочное среднее) и std (среднеквадратичное отклонение) ошибки.

На графике (2) изображены кривые ощибки проведения сегментов для разных моделей. Точка (x, y) на кривой показывает, что данная модель для $y\%$ клиентов дает ошибку сегментации больше, чем x.

Из таблицы (1) и графика (2) видно, что лучшее качество у модели one-hot, незначительно меньшее качество у LDA и $PLSA\_45$. Так же видно, что наиболее сложная (ARTM) модель уступает по качеству более простым (даже Random модели), что может свидетельствовать о потере ею части полезной для сегментирования, но не для определения тематики, информации при переходе к тематическому представлению.

Заметим, что данные таблицы (1) и графика (2) подтверждают гипотезу незначительного ухудшении качества сегментации при переходе

к тематическим векторным представлениям истории транзакций для искусственных профилей.

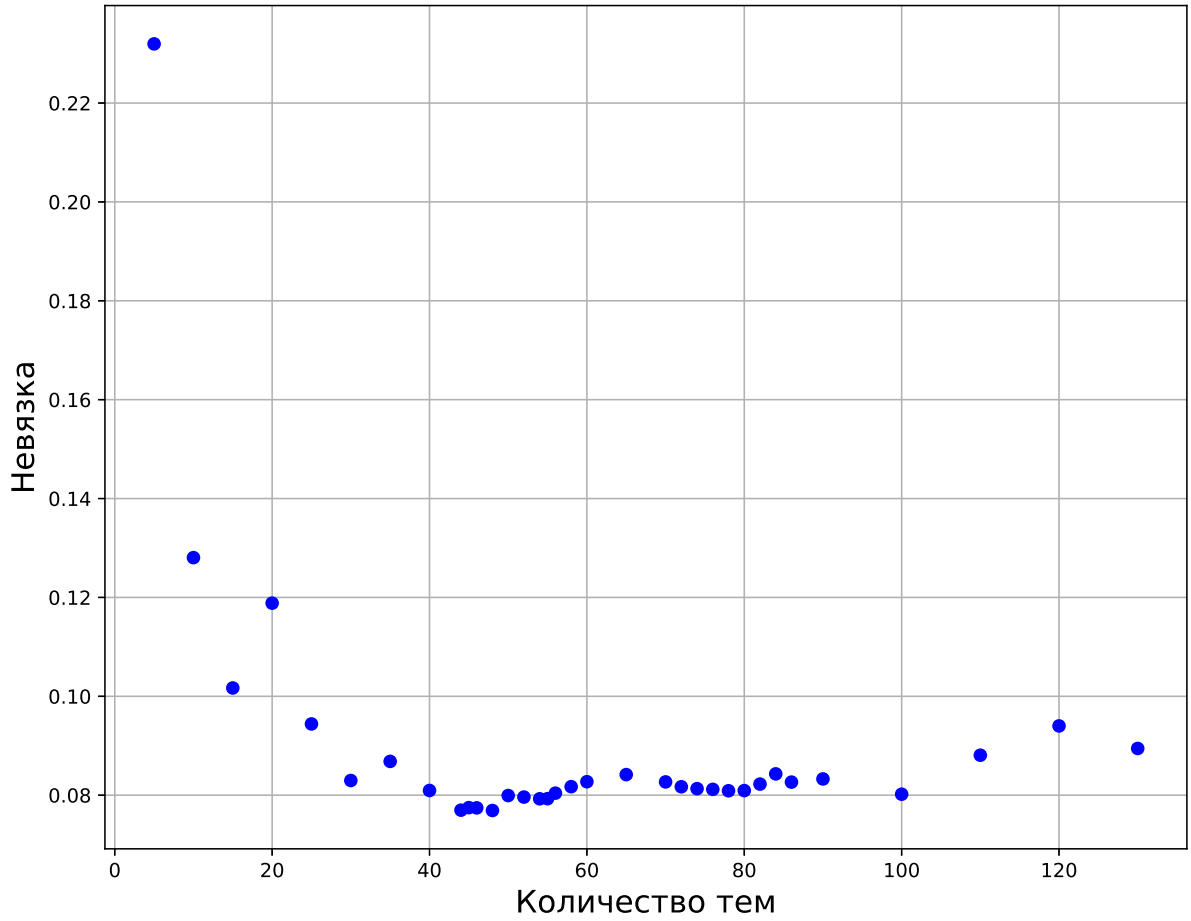

Рис. 3: Зависимость средней ошибки на PLSA от числа тем модели на искусственных профилях

Зависимость качества сегментации от числа тем в модели PLSA

На графике (3) показана зависимость ошибки сегментации с помощью тематических представлений, взятых из модели PLSA, с различным количеством тем. Видно, что модели с количеством тем до 30 значительно уступают по качеству сегментации моделям с большим количеством тем, а наилучшее качество дают модели с количеством тем от 44 до 48.

3.2.2 Сравнение простого и тематического сегментирования

В данном эксперименте измеряется качество моделей с помощью сравнения с простой сегментацией. Сравнивается сегментирование, проведенное с помощью тематического векторного представления (предсказанная

сегментация), и простое сегментирование моделью one-hot (истинная сегментация). В качестве данных для эксперимента используются реальные, а не искусственные профили.

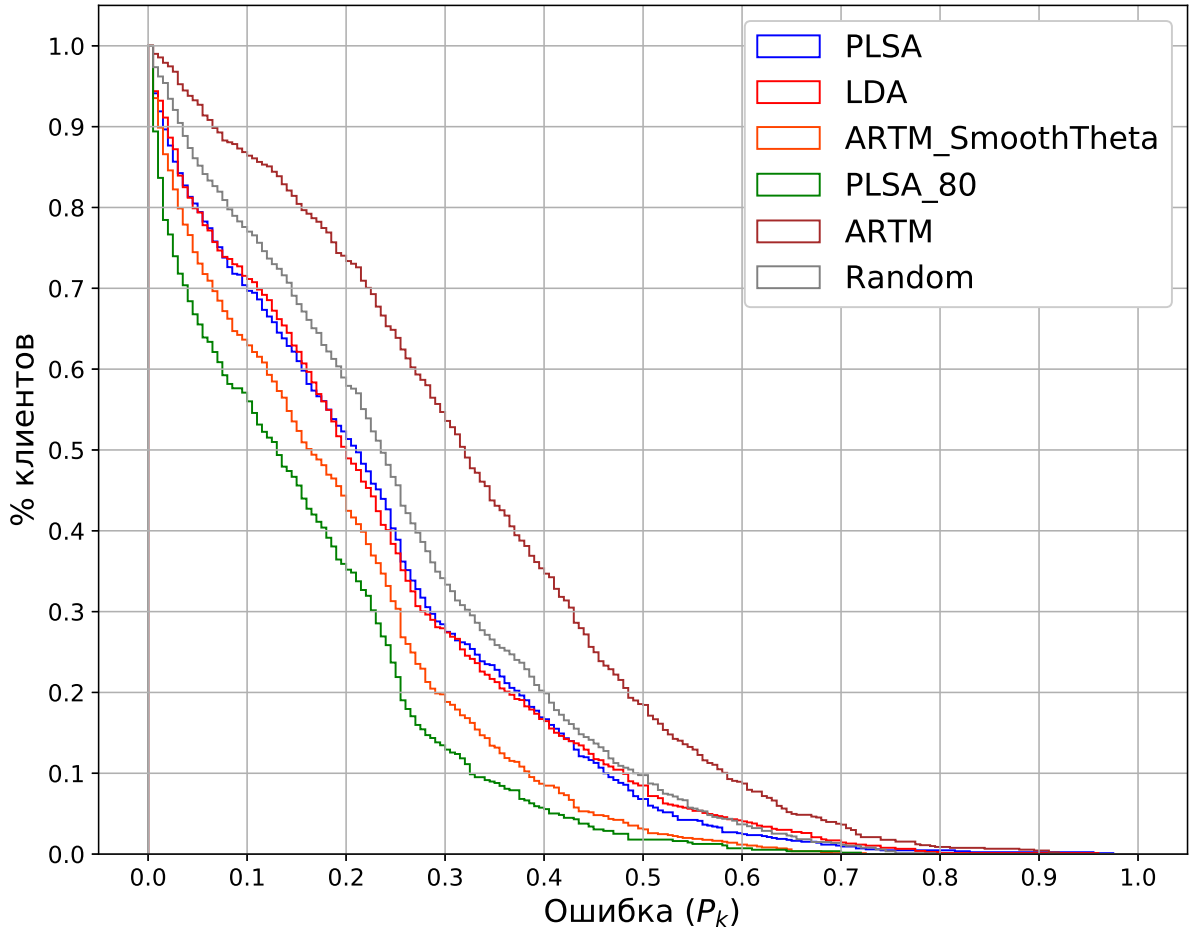

Рис. 4: Сравнение тематического сегментирования с простым на 2000 реальных профилях

Сравнение моделей тематических представлений

Сравниваются границы сегментов проведенных простым и тематическим сегментированием, полученным с помощью различных тематических моделей. $WindowDiff$ и P_k меры между one-hot и тематическим сегментированием занесены в таблицу (2) и отображены на графике (4).

Видим, что наиболее похожим на one-hot по метрике P_k является сегментирование с помощью модели PLSA, построенной на 80 темах (количество тем выбрано оптимизацией P_k ошибки при сравнении с one-hot). Незначительно уступает по качеству перед PLSA модель ARTM_SmoothTheta с коэффициентом регуляризации $\tau = 10^4$,(коэффициент τ получен так же оптимизацией по ошибке P_k при

	P_k		WindowDiff	
Модель сегментации	mean	std	mean	std
lazy	0.564	0.219	0.564	0.219
random	0.237	0.165	0.360	0.193
LDA	0.222	0.176	0.333	0.207
PLSA	0.212	0.169	0.321	0.203
PLSA $p(t \mid d, w)$	0.212	0.170	0.322	0.204
PLSA_45	0.191	0.159	0.291	0.196
PLSA_80	**0.150**	**0.138**	**0.235**	**0.174**
PLSA_80 $p(t \mid d, w)$	0.152	0.144	0.238	0.177
ARTM_SmoothTheta	0.179	0.147	0.285	0.188
ARTM_SmoothTheta $p(t \mid d, w)$	0.170	0.147	0.279	0.187
ARTM	0.327	0.186	0.451	0.205
ARTM $p(t \mid d, w)$	0.286	0.181	0.412	0.211

Таблица 2: Сравнение тематического сегментирования с простым на 2000 реальных профилей

сравнении с one-hot). Значительно менее похоже на one-hot сегментирование с помощью PLSA и LDA.

Значит, с помощью правильного подбора числа тем и коэффициентов регуляризаторов можно значительно улучшить сходство с простым сегментированием.

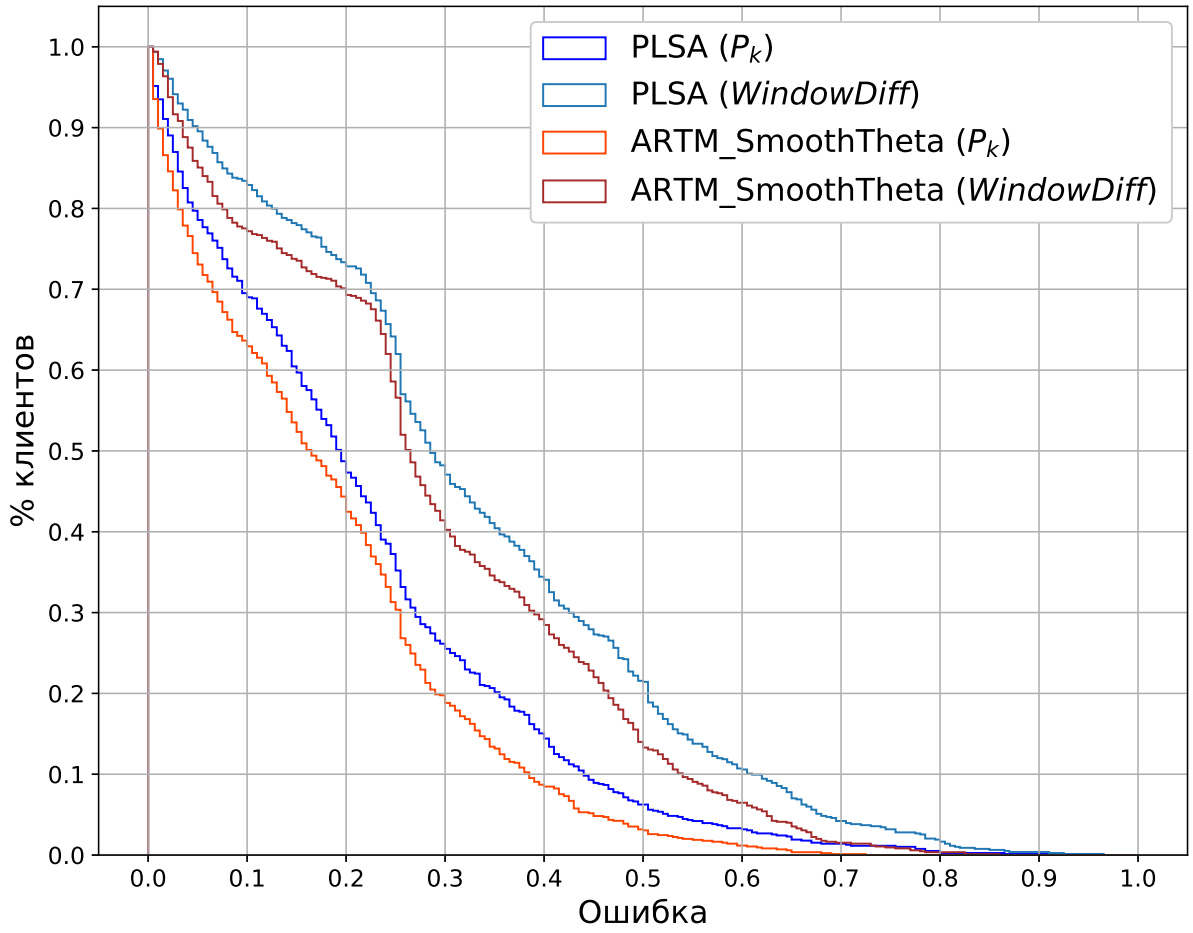

Рис. 5: Сравнение P_k и $WindowDiff$ меры измерения качества

Сравнение P_k и WindowDiff

На графике (5) сравниваются P_k и WindowDiff метрики измерения качества для моделей PLSA и ARTM_SmoothTheta. Из него и таблицы (1) видно, что P_k и WindowDiff метрики одинаково ранжируют модели по качеству проведения сегментов. Поэтому в остальных экспериментах выводятся графики, отображающие только P_k ошибку.

Сравнение тематики транзакции $p(t|d,w)$ и тематики mcc-кода $p(t|w)$

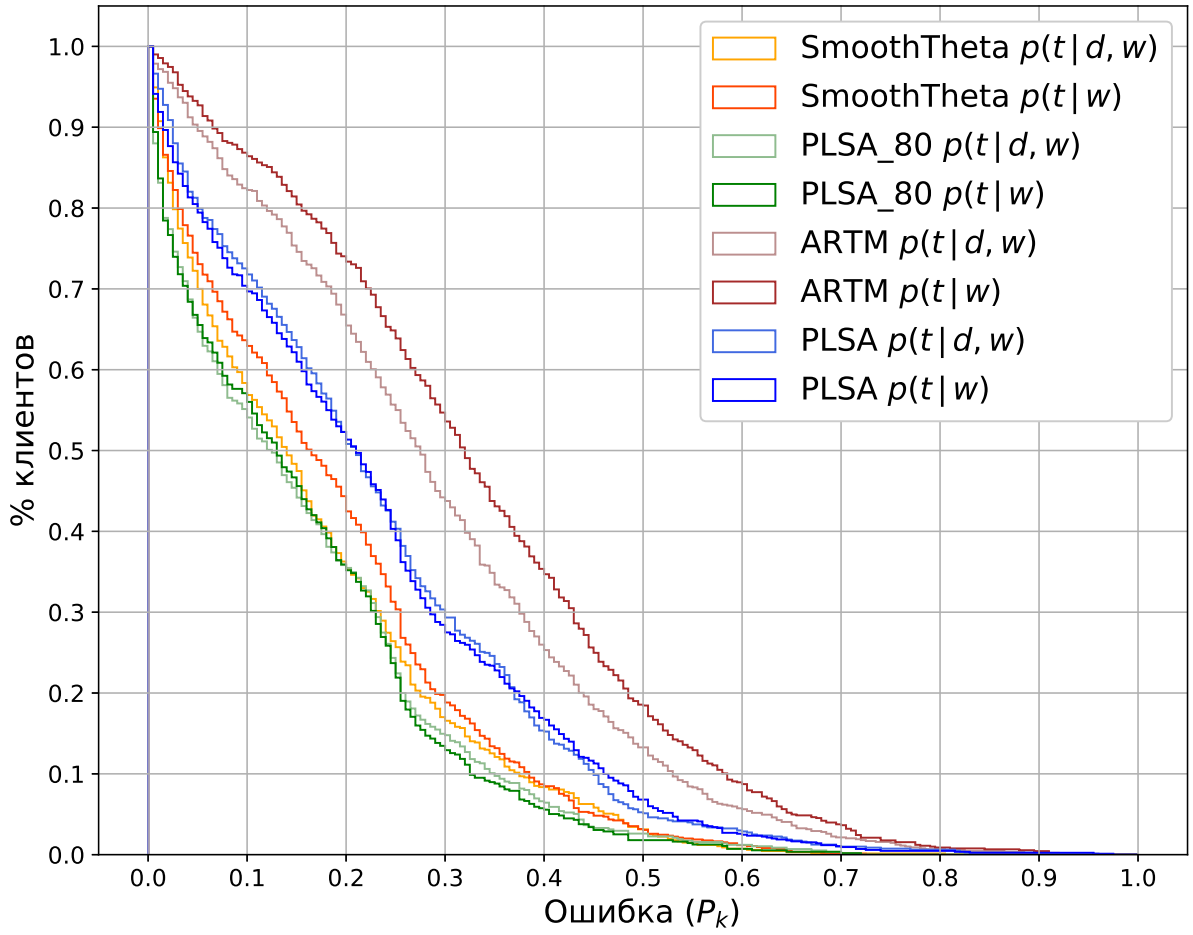

Рис. 6: Сравнение тематических представлений транзакций с помощью их тематик и тематик их mcc-кодов для различных моделей

Из графика (6) и таблицы (2) видно, что для PLSA моделей значительной разницы между $p(t\,|\,w)$ и $p(t\,|\,d,w)$ представлениями не наблюдается. Для моделей с регуляризаторами разница видна. Причем при переходе от тематик mcc-кодов $p(t\,|\,w)$ к тематикам транзакций $p(t\,|\,d,w)$ качество улучшается

Из графиков (6), (4) и таблицы (2) видно, что из рассмотренных моделей наилучшим качеством (по критерию сравнения с простой сегментацией) обладает модель PLSA_80, использующая тематику mcc-кода $p(t\,|\,w)$.

3.3 Демонстрация сегментации истории транзакций

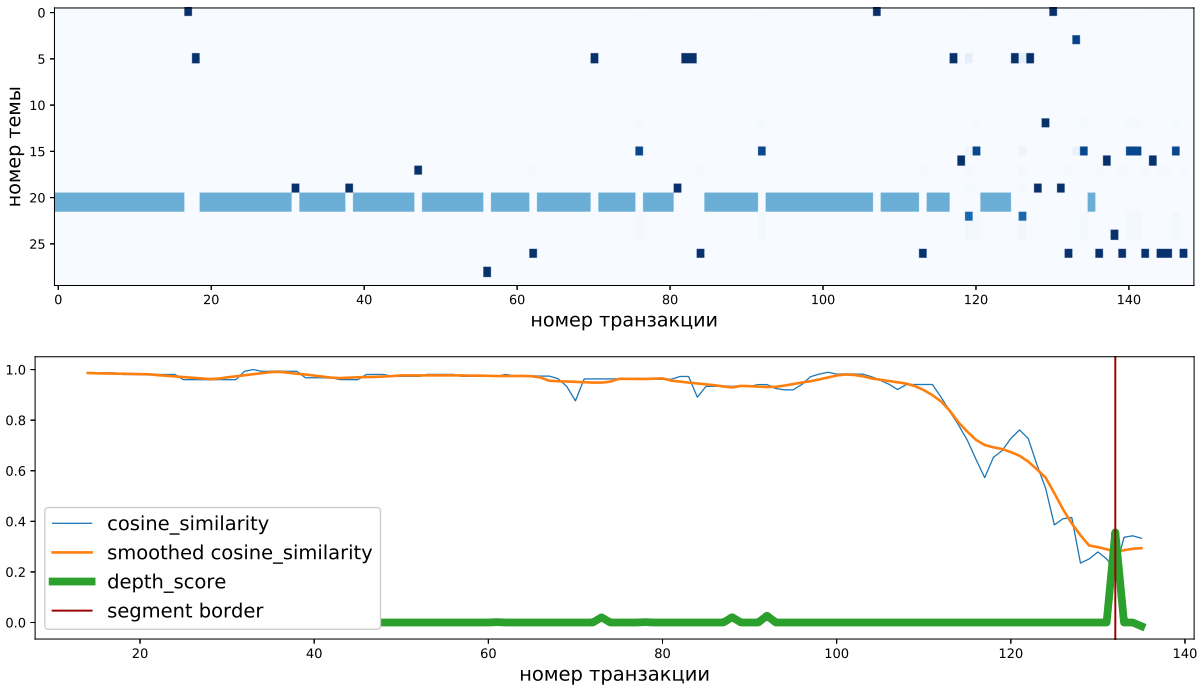

Рис. 7: Результат сегментации истории транзакций клиента

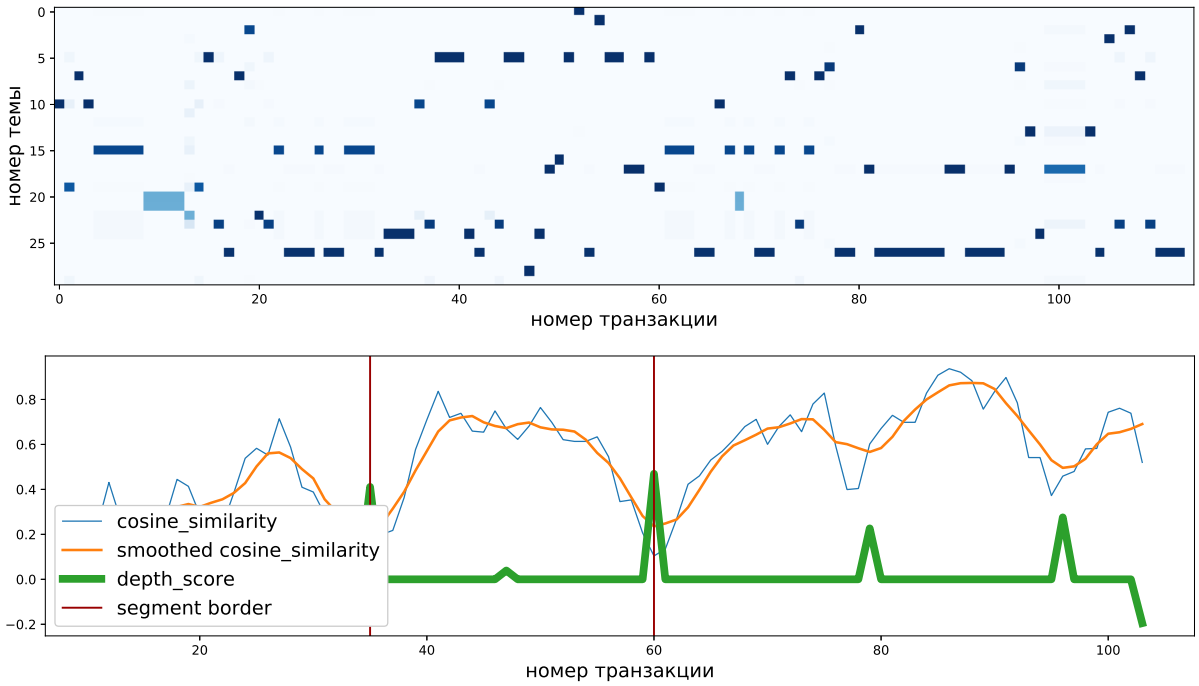

Рис. 8: Результат сегментации истории транзакций клиента

На рисунке (7) показан пример сегментации истории транзакций клиента. Можно предположить, что на протяжении большей части истории

пользования картой клиент использовал её в основном у одного продавца. В какой-то момент клиент расширил свое множество сценариев использования карты и алгоритм, обнаружив этот момент, поставил границу сегмента. На рисунке (8) алгоритм выделил интервал с множеством mcc-кодов, связанных с ремонтом.

4 Заключение

В работе были предложены способы представления истории пользователя с помощью тематик его транзакций и тематик mcc-кодов для последующей сегментации. Даны рекомендации по числу тем и регуляризации для построения тематической модели, служащей для получения тематик, улучшающих качество сегментирования.

Предложены методы оценивания качества модели сегментации транзакционных данных розничных клиентов с помощью искусственных профилей и с помощью сравнения с простой сегментацией.

Подтверждена гипотеза о незначительном изменении качества сегментации при переходе к тематическому представлению, для искусственных профилей.

Продемонстрированы примеры сегментации историй реальных клиентов, представленных в тематическом виде, и предложены их интерпретации.

Список литературы

[1] Никитин Ф.А. Применение мультимодальных тематических моделей к анализу транзакционных данных банков. *Квалификационная работа на соискание степени бакалавра*, 2018.

[2] Thomas Hofmann. Hofmann, t.: Unsupervised learning by probabilistic latent semantic analysis. machine learning 42(1-2), 177-196. 42:177–196, 01 2001.

[3] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, March 2003.

[4] K. V. Vorontsov. Additive regularization for topic models of text collections. *Doklady Mathematics*, 89(3):301–304, May 2014.

[5] Смирнов Е.А. Тематическая сегментация диалогов контактного центра. *Магистерская диссертация*, 2018.

[6] Martin Riedl and Chris Biemann. Text Segmentation with Topic Models . *Journal for Language Technology and Computational Linguistics (JLCL)*, 27(47-69):13–24, 2012.

[7] Arthur U. Asuncion, Max Welling, Padhraic Smyth, and Yee Whye Teh. On smoothing and inference for topic models. *CoRR*, abs/1205.2662, 2009.

[8] Doug Beeferman, Adam L. Berger, and John D. Lafferty. Text segmentation using exponential models. *CoRR*, cmp-lg/9706016, 1997.

[9] Lev Pevzner and Marti A. Hearst. A critique and improvement of an evaluation metric for text segmentation. *Comput. Linguist.*, 28(1):19–36, March 2002.