Исследование способов оценки качества тематических моделей на банковских транзакциях

Василий Алексеев

МФТИ, ИАД
Отчёт по НИР

Осень 2018

Содержание

С чем работаем: примеры транзакций

	client_id_way4	cardnumber	trans_date	amount_rur	trans_crncy
0	158860021	77090664	2014-01-01	-50.0	RUR
1	65191298	193075454	2014-01-01	-50.0	RUR
2	68478688	169746069	2014-01-01	-2000.0	RUR

- `client_id_way4` — идентификатор пользователя
- `cardnumber` — номер карты
- `trans_date` — дата проведения транзакции
- `amount_rur` — сумма транзакции
- `trans_crncy` — валюта

Постановка задачи

- U — множество пользователей (Users)
- $V \subset \mathbb{R}$ — денежные суммы (Values)
- C — множество MCC кодов (Codes)
- $A \subseteq V \times C$ — виды покупательской активности (Actions)
- M — моменты времени (Moments)
- $H(u) \subseteq M \times A$ — история покупательской активности $u \in U$
- $H(u) \ni h = (m, v, c)$ — транзакция

Считаем, что $\exists \mathcal{T}$ — множество покупательских профилей, $|\mathcal{T}| \ll |U|$ — и отображение $f^* : U \to 2^{\mathcal{T}}$, такие что

- $|f^*(u)| \ll |\mathcal{T}| \quad \forall u \in U$
- $|f^{*-1}(t)| \ll |U| \quad \forall t \in \mathcal{T}$
- $f^*(u_i) \cap f^*(u_j) \neq \varnothing \Leftrightarrow$ пользователи u_i и u_j *похожи*

ДНК

Дано

- пользователи U
- истории их покупок $H(u)\ \forall u \in U$

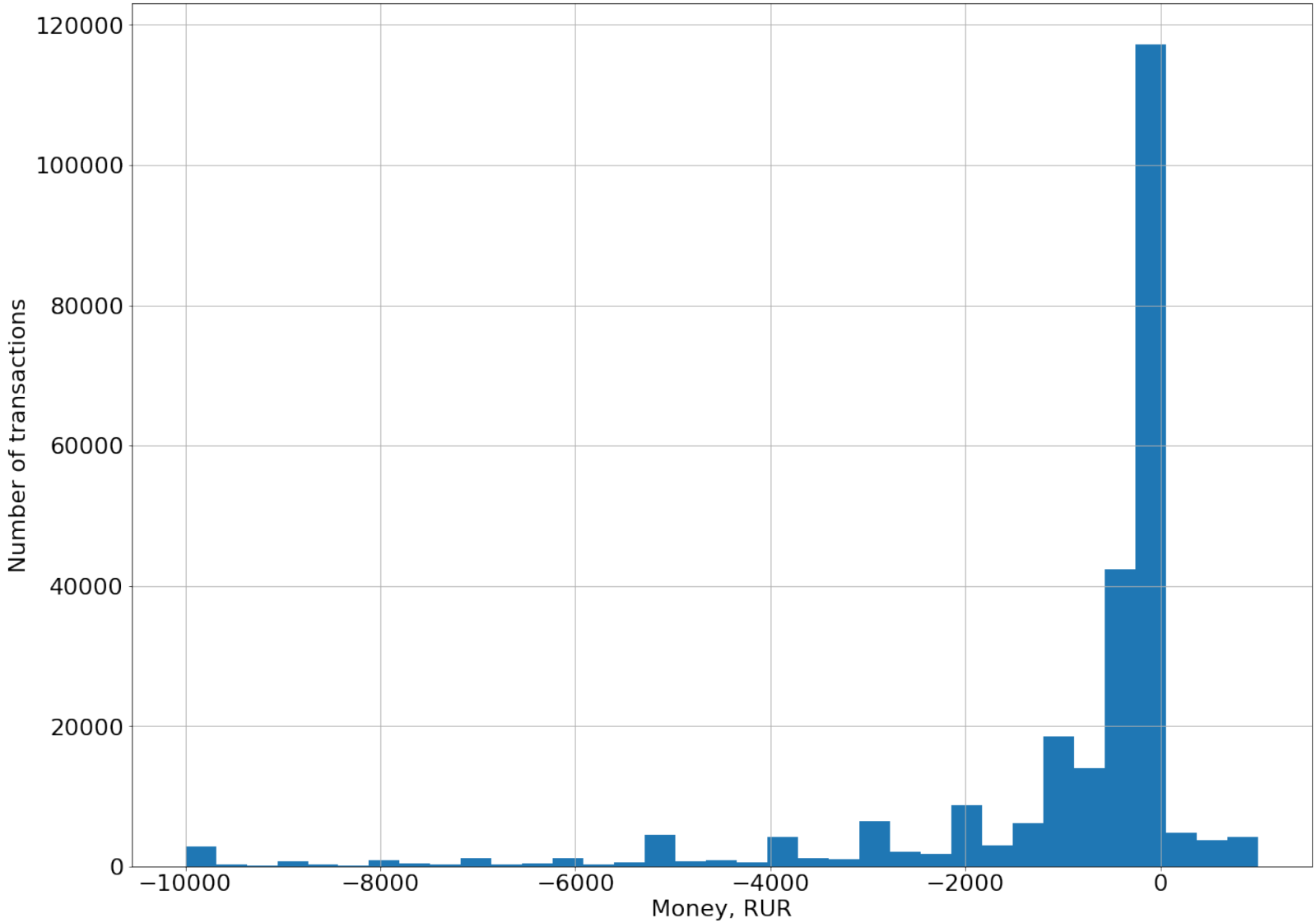
Найти

- интерпретируемое множество профилей $\mathcal{T}$
- соответствующую функцию $f^* : U \to 2^{\mathcal{T}}$

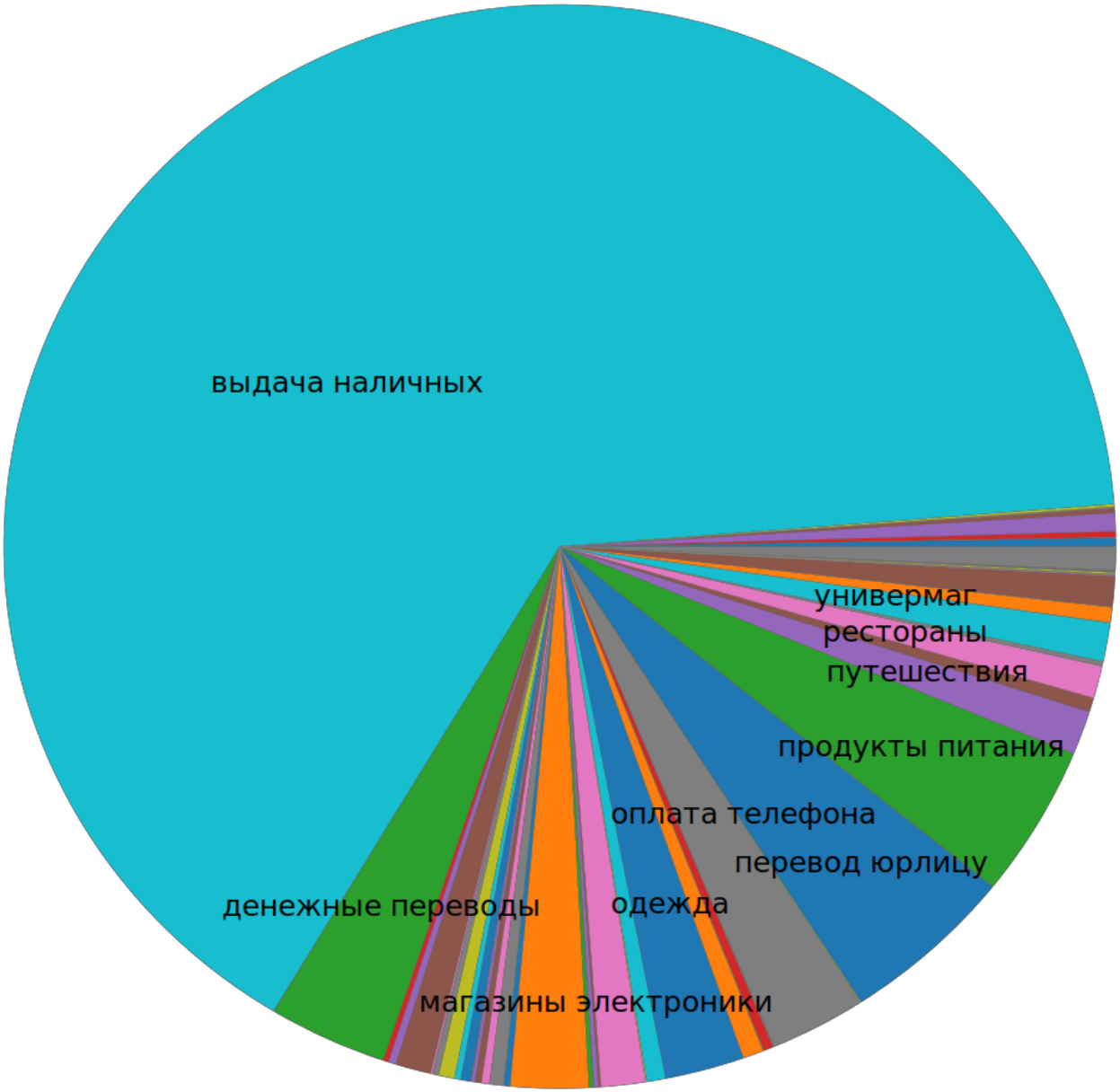
Данные

- 92682 пользователей
- 300000 транзакций
- 77 категорий MCC кодов, 244 MCC кодов
- разные валюты: 200000 транзакций в рублях, ...
- деньги могут тратиться/начисляться
- 50000 из 300000 транзакций — с невалидными MCC кодами и валютами
- временной интервал 4 месяца: от января до апреля 2014 года

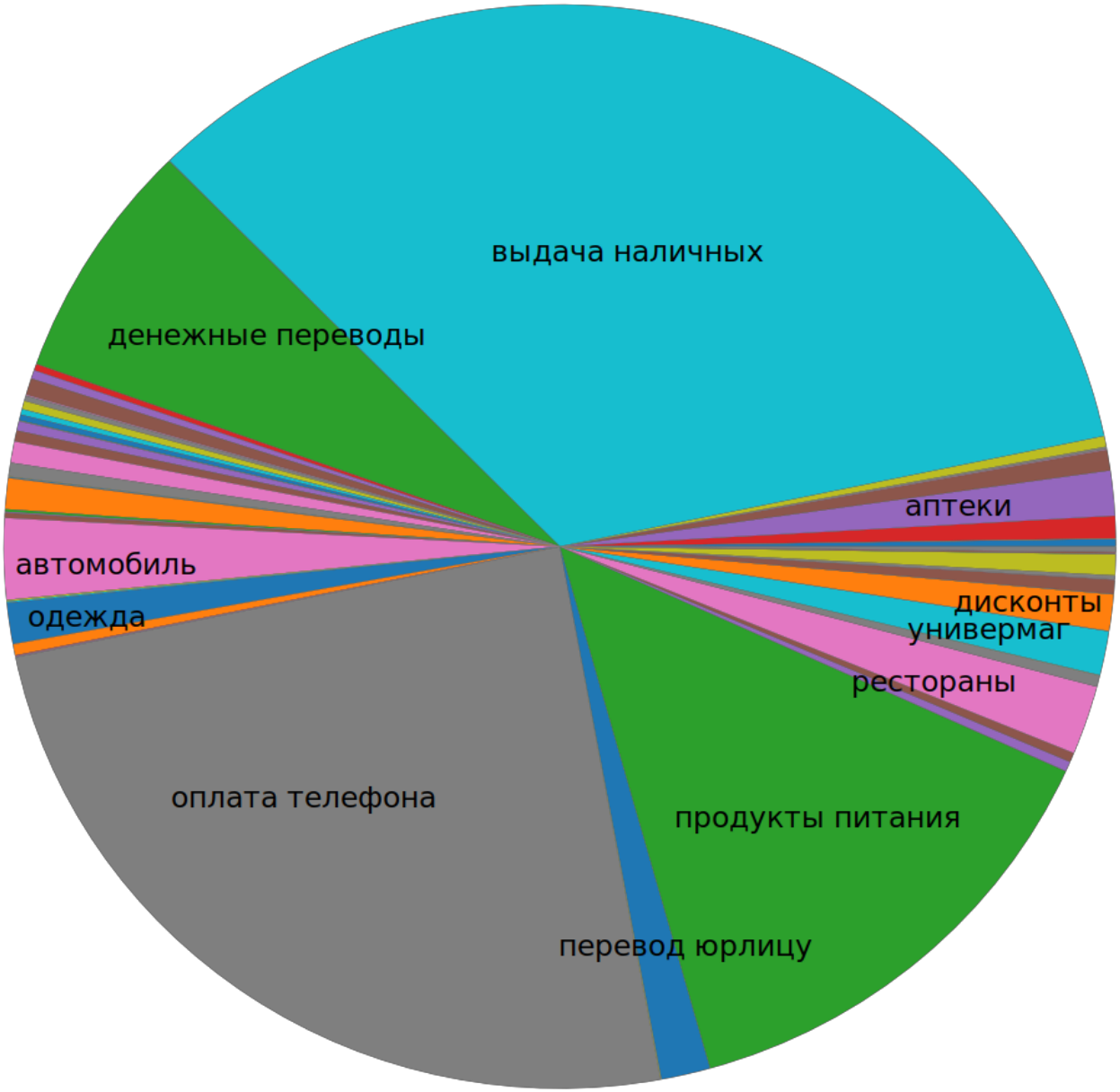
Данные. Распределение трат пользователей



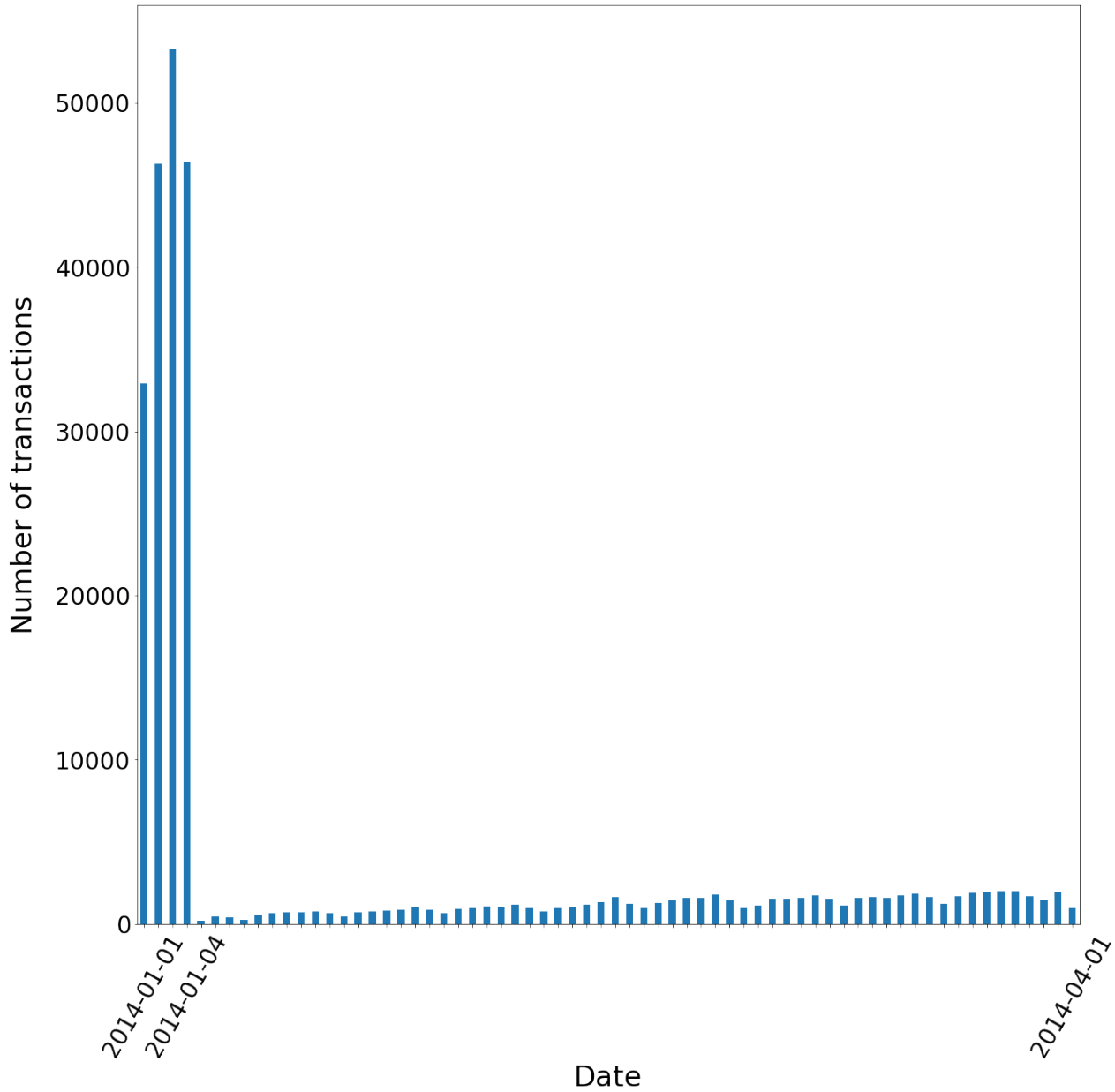
Данные. Потраченные деньги по MCC категориям



Данные. Количество транзакций по MCC категориям



Данные. Транзакции по дням



Содержание

Тематическая модель

Аналогия с тематическим моделированием

- Документ — история транзакций пользователя $H(U) \leftrightarrow D$
- Слова — возможные действия внутри транзакции $A \leftrightarrow W$
- $p(a \mid u) = \sum_{t \in T} p(a \mid t) p(t \mid u)$
- $\Phi \equiv \big(p(a \mid t)\big)_{A \times T}$ $\Theta \equiv \big(p(t \mid u)\big)_{T \times U}$

Но

$|A| = |\mathbb{R} \times C| = |\mathbb{R}|$, поэтому для того, чтобы использовать тематические модели, надо будет перейти от A к $\widetilde{A}$, $|\widetilde{A}| < \infty$

Модальности

$\widetilde{A} = C$

- Сколько раз c_i встречается в истории $H(u)$
- Суммарные траты по c_i для u

$\widetilde{A} = C \times C$

Совстречаемости c_i, c_j в окне

- по времени: $|t_i - t_j|$
- по числу транзакций: $\left|\left\{(t, m, c) \in H(u) : t \in [t_i, t_j]\right\}\right|$

$\widetilde{A} = \widetilde{V} \times C$

Категории трат $\widetilde{V} \leftarrow V \subset \mathbb{R}$, $|\widetilde{V}| < \infty$ по

- всем транзакциям всех пользователей
- транзакциям с данным c_i всех пользователей
- всем транзакциям данного пользователя

Модальность: количество c_i

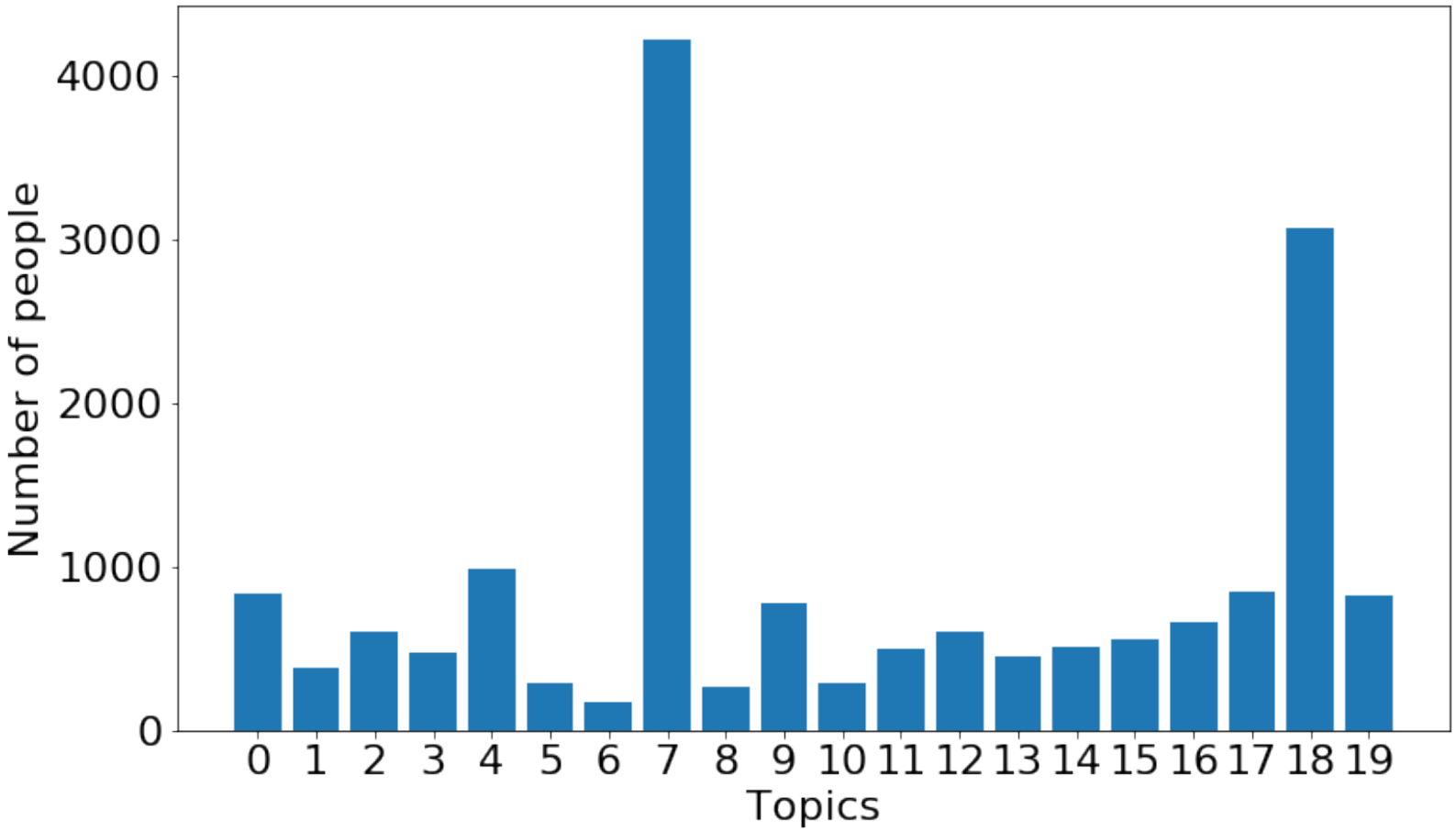
Примеры тем (по количеству MCC категорий)

- книги, жкх, путешествия
- дом, путешествия, активный отдых
- дети, аптеки, загородный дом

Примеры тем (по количеству MCC кодов)

- продукты питания, дом, кинотеатры
- интернет-продажи, ремонт
  25% молодых, 60% средних лет, 15% пожилых
- аптеки, косметология, ветеринары и зоотовары
  15% молодых, 25% средних лет, 60% пожилых

Модальность: количество c_i (MCC категория)



$topic_7$ = "денежные переводы"; $topic_{18}$ = "продукты питания"

Модальность: суммарные траты по c_i

Примеры тем (при учёте знака при числе траты)

- автомобиль (-), азартные игры (+), авиа (+)
- дом (-), ремонт (-), ночные клубы (-)

Примеры тем (при учёте валюты)

- одежда (RUR), авиа (RUR), дьютифри (RUR), путешествия (SEK), универмаги (EUR)
- косметология (RUR), одежда (RUR), интернет продажи (EUR), услуги интернета (RUR), b2b магазины (RUR) 65% молодых, 30% средних лет, 5% пожилых
- продукты питания (RUR), ресторары (RUR), косметология (RUR), цветочные магазины (RUR) 45% жен, 55% муж
- дом (RUR), путешествия (THB), путешествия (TRY), одежд (THB), косметология (THB)

Модальность: совстречаемости c_i, c_j

Примеры тем (окно в 5 транзакций)

- алкоголь/рестораны, универмаги/одежда, одежда/дом
- рестораны/ночные клубы, ночные клубы/рестораны

Примеры тем (окно в 1 день)

- универмаги/дисконты, рестораны/косметология, автомобиль/косметика
  60% муж, 40% жен

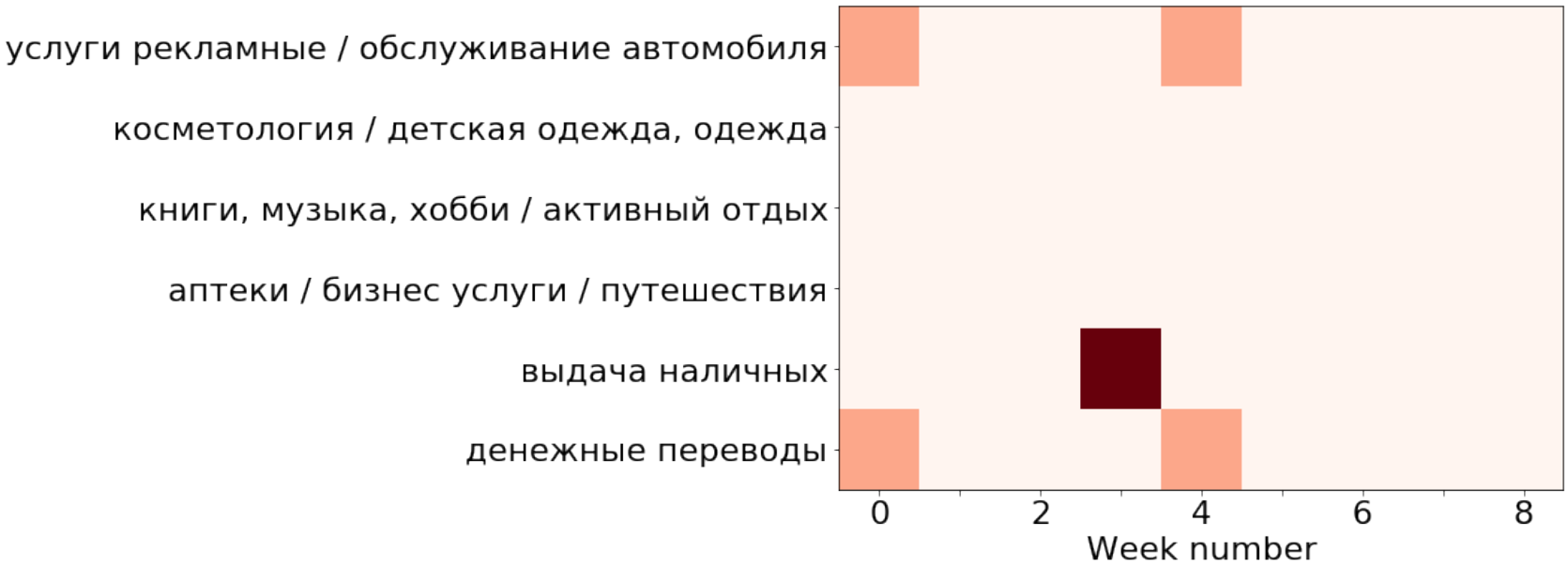
Модальность: категории трат

Примеры тем (тратам на u_i)

- видеоигры (medium), активный отдых (medium), видеоигры (low), одежда (medium), путешествия (high)
  20% молодых, 65% средних лет, 15% пожилых
- услуги рекламные (low), книги и хобби (low), азартные игры (low), книги (medium), азартные игры (relatively-high)
  5% молодых, 35% средних лет, 60% пожилых
- аптеки (medium), рестораны (relatively-high), продукты питания (relatively-high), услуги рекламные (medium), аптеки (high)
  10% молодых, 25% средних лет, 65% пожилых

Изменение профилей пользователей со временем

- Обучение на данных без выделенной группы пользователей
- Разбиение временного интервала по неделям
- Построение Θ для этих пользователей

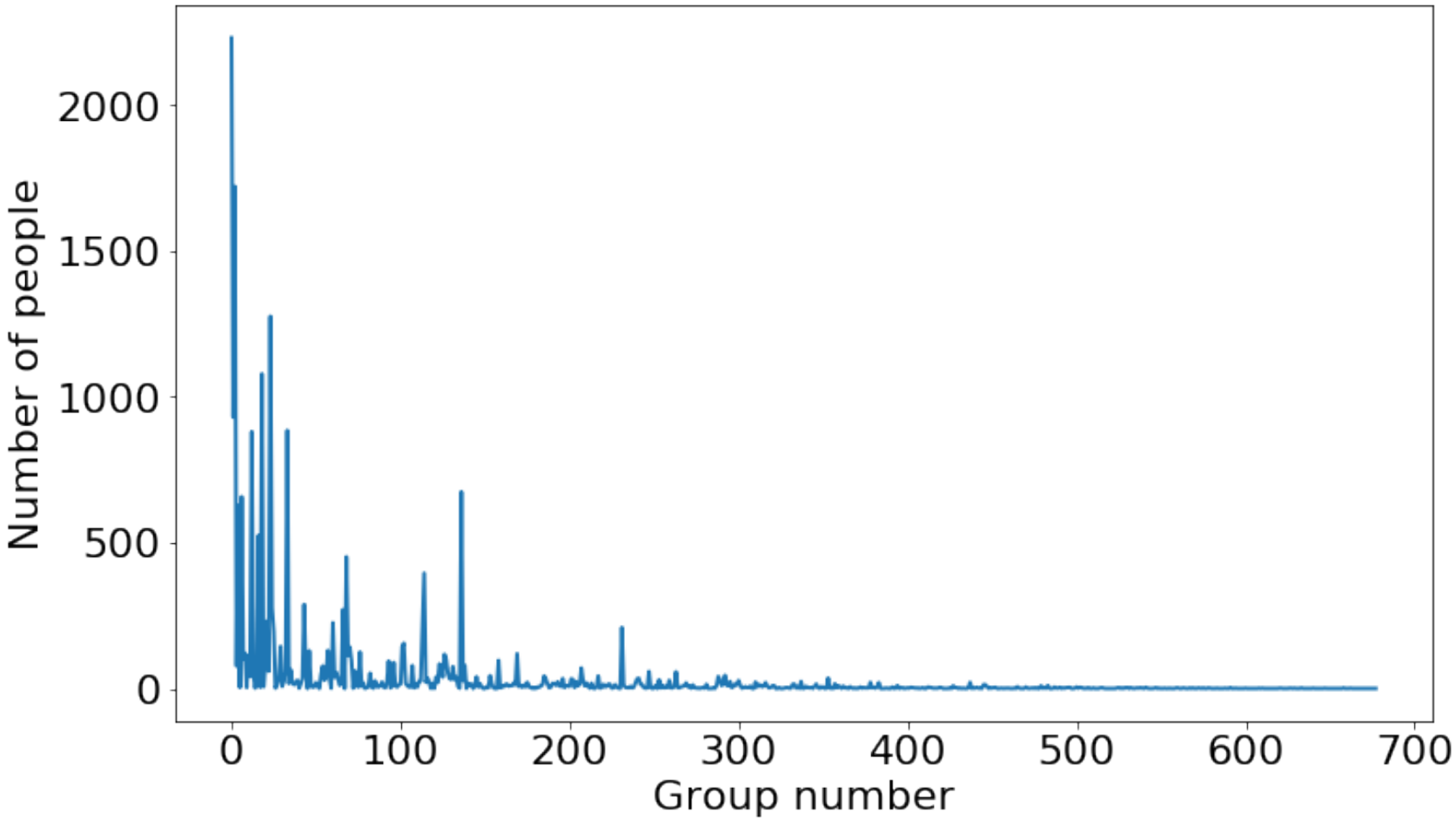


Иерархическая + тематическая

- $u \mapsto \boldsymbol{v} = \boldsymbol{counts}_c \odot \boldsymbol{money}_c$
- 2 топ-компоненты $\boldsymbol{v}$ образуют имя группы
- Объединяем u с одинаковыми топ-компонентами $\boldsymbol{v}$
- Внутри каждой группы — тематическое моделирование

Иерархическая по MCC категориям

- 700 групп, 90-ая персентиль 40 человек в группе
- Оставляем 10% самых населённых. Медиана 100 человек
- В каждой группе — тематическая модель на 4 темах



Примеры групп

- книги и хобби / автомобиль
  - автобобиль (топливо), автобобиль (запчасти)
- одежда / рестораны
  - закусочные и рестораны, семейная одежда, аксессуары
  - мужская и женская одежда, закусочные и рестораны
- кинотеатры, театры, цирки / рестораны
  - фаст фуд, кино, театры
  - закусочные и рестораны, кино
- косметология / одежда
  - косметология, аксессуары
  - мужская и женская одежда, продуктовые магазины
  - магазины семеной одежды, денежные переводы

Тематическая наоборот

Идея

По транзакции понять, какой пользователь мог её совершить

$$p(u \mid c) = \sum_{t'} p(u \mid t')p(t' \mid c)$$

$$\Phi' = (p(u \mid t')), \Theta' = (p(t' \mid c))$$

Тематическая наоборот. Переход $\Phi', \Theta' \to \Phi, \Theta$

$$p(c \mid u) = \frac{p(u \mid c)p(c)}{\sum_c p(u \mid c)p(c)}$$

$p(c)$ — частотная оценка по транзакциям пользователей

$$\begin{cases} \theta_{tu} = p(t \mid u) = \dfrac{p(u \mid t)p(t)}{\sum_t p(u \mid t)p(t)} = \dfrac{\varphi'_{ut}p(t)}{\sum_t \varphi'_{ut}p(t)} \\ p(t) = \sum\limits_c p(t \mid c)p(c) = \sum\limits_c \theta'_{tc}p(c) \end{cases}$$

$$\varphi_{ct} = p(c \mid t) = \frac{p(t \mid c)p(c)}{\sum_c p(t \mid c)p(c)} = \frac{\theta'_{tc}p(c)}{\sum_c \theta'_{tc}p(c)}$$

Примеры тем

- автомобиль, ресторан, алгкогольный магазин, универмаг, книги и хобби
- бизнес услуги, авиа, услуги страхования
- b2b магазины, знакомства
  10% молодых, 20% средних лет, 70% пожилые

Кластеризация графов активностей

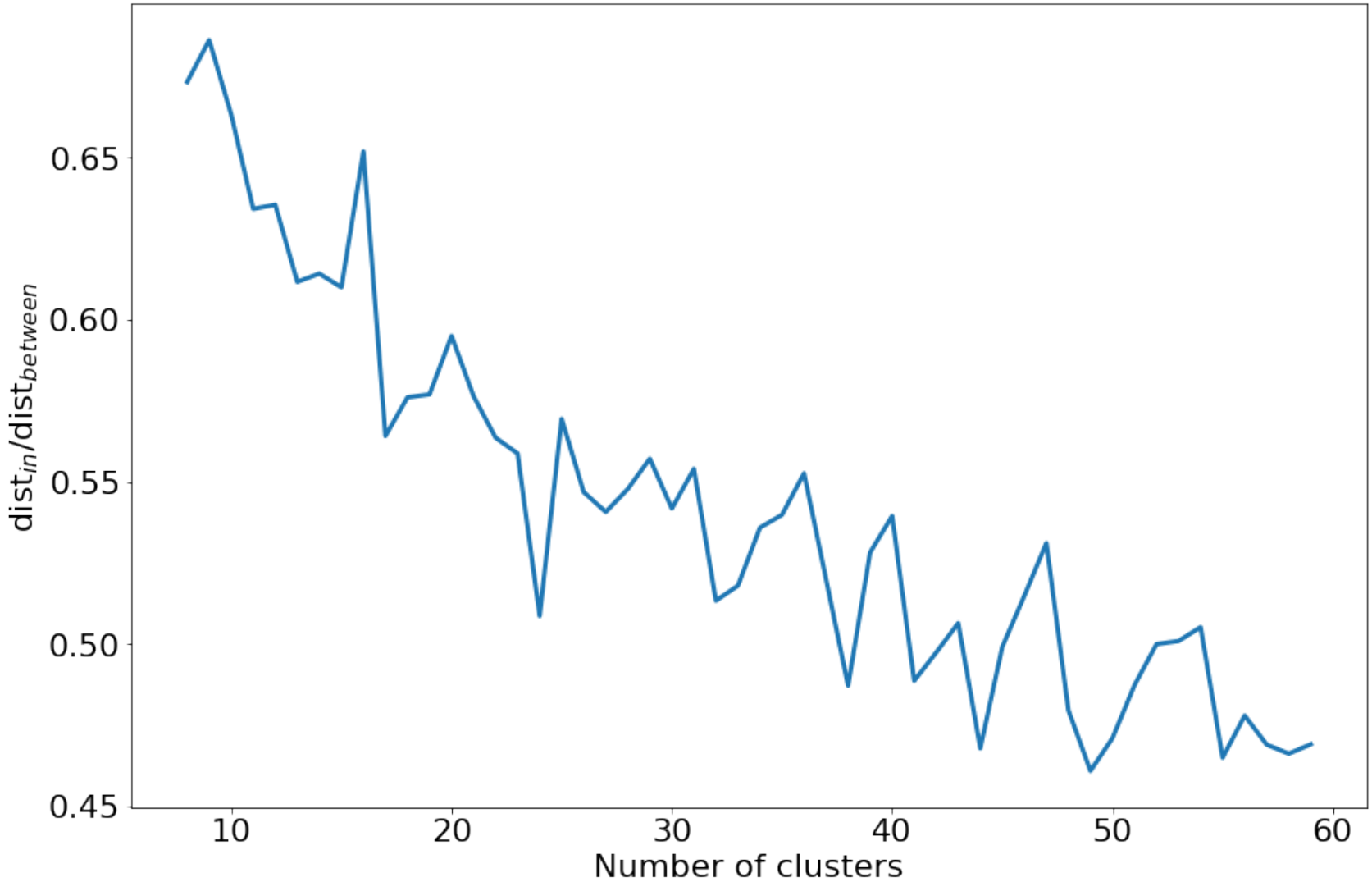
Взвешенный орграф пользовательской активности $g \in G$

$$u \mapsto g \in G$$

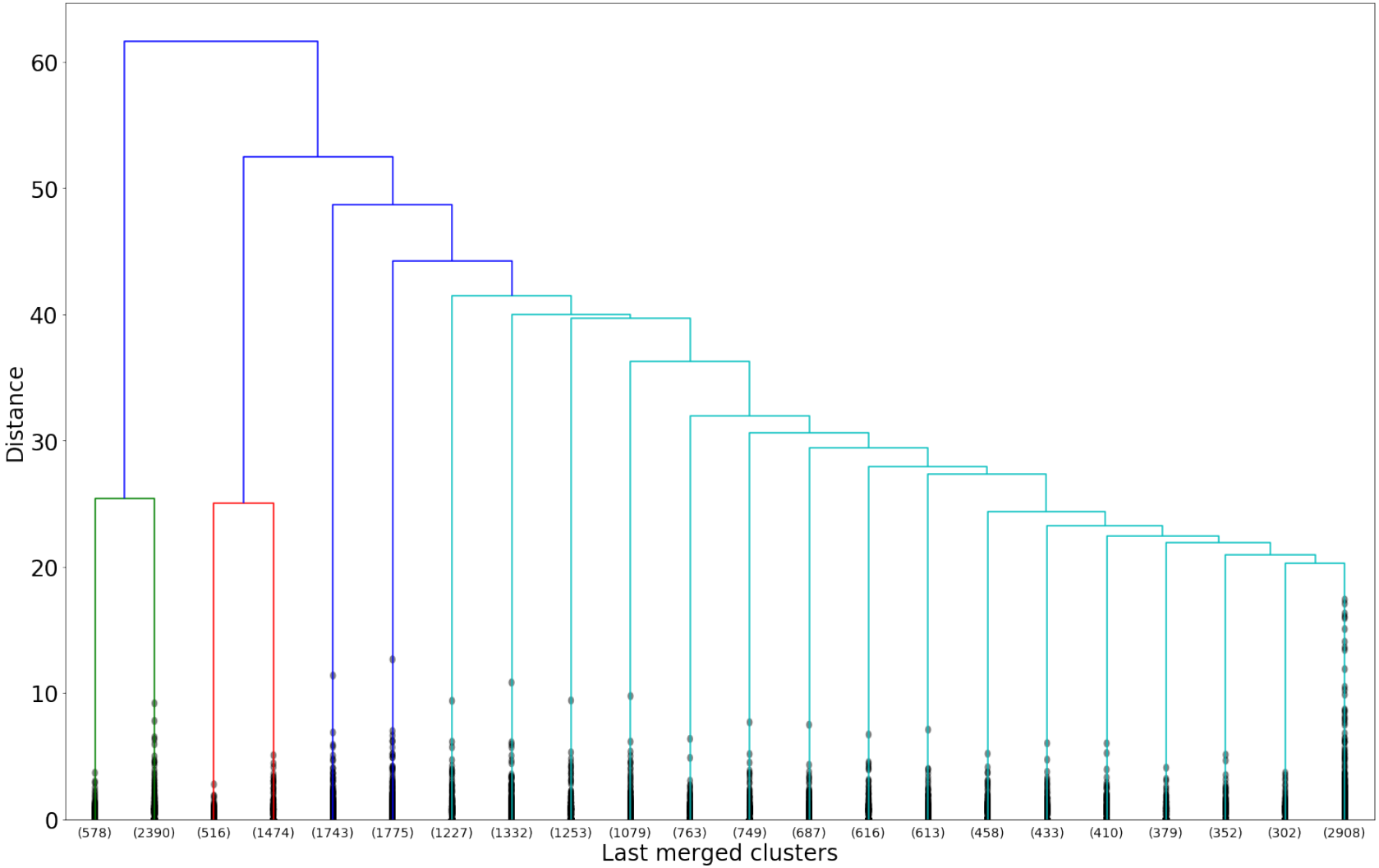
- $root$ — фиктивная вершина, из которой начинаются переходы в случае долгого бездействия пользователя u
- $E = \{(root, c_i) \mid c_i \in C\} \cup \{(c_i, c_j) \mid c_i,\ c_j \in C\}$
- $w(c_i, c_j) \quad \propto \#\left(i : \begin{cases} c(h_i) = c_i,\ c(h_{i+1}) = c_j \\ |t(h_i) - t(h_{i+1})| \leqslant \text{threshold} \end{cases}\right)$
- $w(root, c_i) \propto \#\left(i : \begin{cases} c(h_i) = c_i \\ |t(h_i) - t(h_{i-1})| > \text{threshold} \end{cases}\right)$

$$g \to \boldsymbol{v} = \boldsymbol{w}_e$$

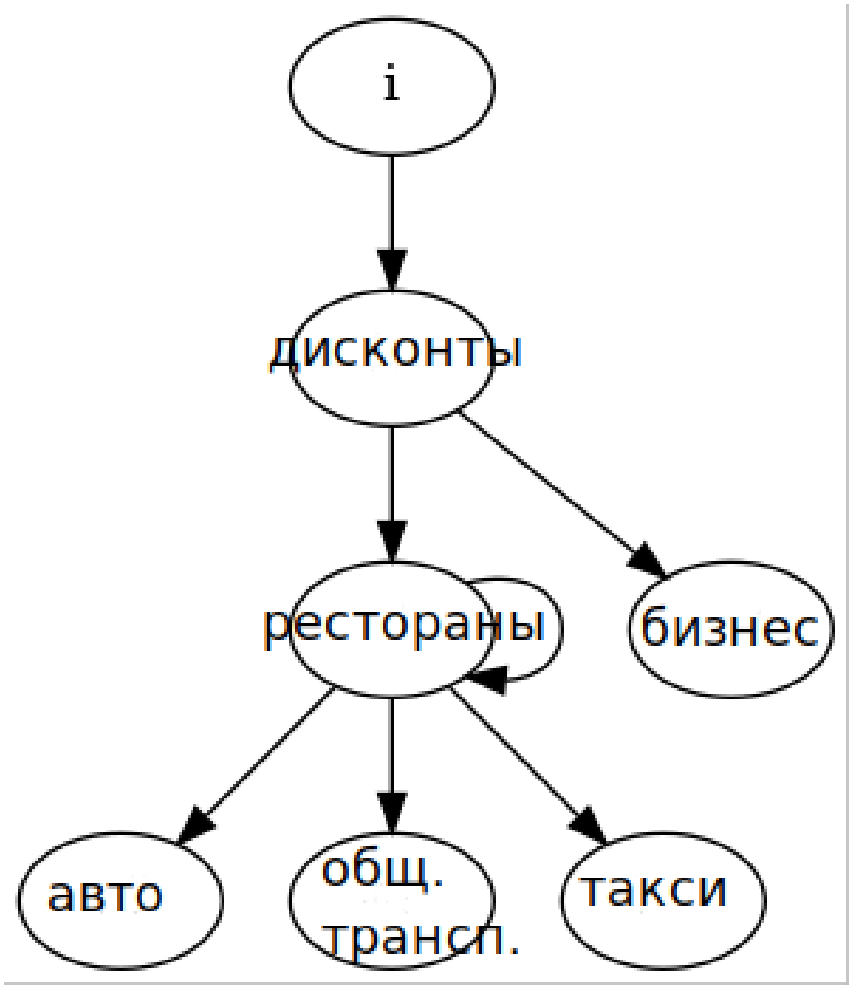
Кластеризация орграфов пользователей. KMeans



Кластеризация орграфов пользователей. Иерархическая



Пример графа



Содержание

Результаты

- Модальности
- Темы
- Функция качества
- Визуализация