Московский государственный университет имени М.В. Ломоносова

Факультет Вычислительной Математики и Кибернетики

Кафедра Математических Методов Прогнозирования

Попов Николай Олегович

Гиперграфовые тематические модели транзакционных данных

МАГИСТЕРСКАЯ ДИССЕРТАЦИЯ

Научный руководитель:

академик РАН

К. В. Рудаков

Москва, 2018

Содержание

1 Введение

В настоящее время деятельность человека порождает огромные массивы сложно структурированной информации, такие как информация в социальных сетях, рекомендательных систем, информация о товарах и услугах и т. д. Наиболее ярким примером подобных данных являются данные о банковских транзакциях между компаниями (юридическими лицами), так как данные о банковских транзакциях в прямом смысле являются продуктом деятельности компании и в полной мере отражают вид экономической деятельности компании. Анализ больших объемов данных о банковских транзакциях позволит банкам улучшить свои услуги, за счет использования новой, ранее не используемой информации. Каждая транзакция может содержать в себе как общие паттерны: такие как этап развития компании, так и более узкие: паттерны в рамках экономической отрасли. Данная информация может принести клиентам банка огромную пользу, с ее помощью клиенты способны формировать профиль контрагентов и конкурентов, что позволит быстрее перенимать новые способы ведения бизнеса.

Анализ транзакционных данных позволит банкам иметь более таргетированные финансовые услуги, а также даст возможность по формированию новых услуг в области отраслевого и финансового консалтинга. Однако, в настоящее время, крупные банки не имеют полноценных инструментов по анализу транзакционных данных, поэтому становится актуальным создание инструментов, которые позволят более подробно изучить финансовые потоки. Транзакционные данные охватывают сотни тысяч различных компаний, при этом похожие компании, например, производящие одинаковые продукты, могут и вовсе не иметь общих контрагентов. Таким образом для формирования профиля клиента, имеет смысл переход от конкретного контрагента компании к более общей сущности – идентификатору отрасли, в которой работает контрагент. Таким образом в данной работе будет рассматриваться транзакции между клиентом и некоторой финансовой отраслью, к которой принадлежит контрагент. В России при регистрации юридического лица, обязательным является указание общероссийского классификатора видов экономической деятельности (ОКВЭД), но к сожалению, компании относятся халатно к его заполнению, в связи с чем возникает задача по восстановлению отрасли экономической деятельности компании. Задача по восстановлению отрасли деятельности компании имеют некоторые решения [6], в данной же работе для предварительной обработки данных использовался алгоритм частичного обучения (semi-supervised) label propagation.

Формирование профиля компании является одной из приоритетных задач банка, так как

подобные признаки универсальны, и могут быть использованы в решении практически любой задачи машинного обучения. В данной работе рассмотрена одна из приоритетных задач – поиск клиентов на этапе развития, так как такие компании являются основной массой клиентов по кредитным продуктам банка. Эксперты выделяют порядка 8 различных видов инвестиционных кредитов: на покупку оборудования, недвижимости, закупку транспортных средств и т. д.

В данной работе ставится задача восстановления латентной информации о видах необходимого кредитования по наблюдаемым транзакционным данным. К сожалению, в реальной жизни отдел продаж имеет ограниченные ресурсы, поэтому при внедрении данного решения появляется побочная, но не менее важная задача – ранжирование клиентов, один из способов решения данной задачи так же предложен в данной работе. Новизна данной работы заключается в том, что математической моделью для представления мультимодальных транзакционных данных является гиперграф. Вершинами гиперграфа являются объекты различных модальностей, причем с каждой вершиной связан неизвестный латентный тематический профиль. Наблюдаются транзакции между объектами, которые описываются рёбрами гиперграфа. Предполагается, что вероятность транзакции определяется степенью сходства тематических профилей входящих в нее объектов.

2 Тематическое моделирование

В данном разделе будет описана классическая постановка задачи тематического моделирования. Имеется два конечных множества: D — коллекция текстовых документов и W — множество употребляемых в них терминов. Имеются наблюдаемые данные о парах «документ–термин» $(d, w) \in D \times W$ — число вхождений n_{dw} термина w в документ d. Предполагается, что существует конечное множество латентных тем T, и что появление каждой пары (d, w) связано с некоторой темой $t \in T$. Другими словами, коллекция документов является случайной выборкой независимых наблюдений из распределения $p(d, w, t)$ на дискретном вероятностном пространстве $D \times W \times T$.

Вероятностная модель порождения данных связывает вероятность $p(d, w)$ совместного появления пар «документ–термин» с вероятностями $p(d, t)$ и $p(w, t)$, которые выражаются через условные вероятности двумя способами:

$$p(d, t) = p(d \mid t)p(t) = p(t \mid d)p(d)$$
$$p(w, t) = p(w \mid t)p(t) = p(t \mid w)p(w).$$

Существуют два эквивалентных варианта записи вероятностной порождающей модели — симметричный

$$p(d, w) = \sum_{t \in T} p(d \mid t)p(w \mid t)p(t),$$

и асимметричный

$$p(d, w) = \sum_{t \in T} p(w \mid t)p(t \mid d)p(d).$$

Оба основаны на формуле полной вероятности и гипотезе условной независимости, которая соответственно допускает две эквивалентные формы записи:

$$p(d, w \mid t) = p(d \mid t)p(w \mid t);$$
$$p(w \mid t, d) = p(w \mid t).$$

Чтобы найти неизвестные распределения по исходным данным, ставится задача максимизации правдоподобия:

$$\sum_{d,w} n_{dw} \ln p(d, w) \to \max.$$

В модели PLSA [7] исходно рассматривались оба варианта, в модели LDA [8] используется только асимметричный вариант. В обоих случаях возникает задача стохастического матричного разложения, приближённого в смысле дивергенции Кульбака–Лейблера.

В коллаборативной фильтрации возникает аналогичная задача, с точностью до замены терминологии: «документы» → «пользователи», «термины» → «предметы», «темы» → «интересы».

Более общая модель предполагает, что документ состоит не только из терминов, но может также содержать токены других типов: фотографии, видеозаписи, пользователи, отрасль экономической деятельности, жанр и т. д. Конечное множество (словарь) токенов одного типа называется *модальностью*. Каждый токен v, какой бы модальности он ни был, связан с латентными темами аналогично терминам:

$$p(v,t) = p(v \mid t)p(t) = p(t \mid v)p(v).$$

Во многих прикладных задачах этого обобщения оказывается недостаточно. Данные могут охватывать сложные взаимодействия между объектами разных типов. В таких случаях объединение модальностей только внутри документов становится неестественным ограничением.

Например, в рекомендательных системах имеется информация о пользователях: они посещают страницы с обзорами, пишут отзывы, переходят по рекламным объявлениям, все эти объекты также состоят из токенов.

В социальных сетях пользователи взаимодействуют друг с другом, размещают у себя на странице фотографии, документы, оценивают записи, состоят в сообществах и т. д.

Компании также взаимодействуют друг с другом по средствам банковских транзакций, банковская транзакция по своей сути тоже является объектом, так как помимо суммы перевода, она содержит много важной мета-информации, наиболее значимой является текстовое описание назначения платежа. Таким образом образуется связь между компанией отправителем денежных средств, компанией получателем, и словами, которые входят в назначение платежа.

3 Тематические модели на гиперграфах

Тематические модели на гиперграфах являются дальнейшим обобщением многомодальных моделей.

Рассмотрим *ориентированный гиперграф* $\Gamma = \langle V, E \rangle$, который определяется множеством вершин V и семейством подмножеств вершин E. Каждое ребро $e \in E$ является линейно упорядоченным подмножеством вершин, $e \subset V$.

Наблюдаемыми данными являются взаимодействия двух или более объектов (транзакции). Таким образом в терминах гиперграфа, объекты являются вершинами, а транзакции – ребрами. В классических тематических моделях вхождение термина w в документ d является транзакцией. Ребро гиперграфа в данном случае соединяет две вершины: $e = (d, w)$, а n_{dw} число вхождений данного ребра в наблюдаемых данных. В реальном мире имеется множество примеров, в которых ребро соединяет более двух вершин.

Например, в социальной сети транзакция «пользователь u лайкнул запись w на странице пользователя v» представляется ребром из трёх вершин $e = (v, u, w)$.

В банковской платежной системе транзакция «слово w содержится в назначении платежа c от компании u к компании v» представляется ребром из трех вершин $e = (w, u, v)$. Данные примеры показывают, что не всегда взаимодействие нескольких вершин можно свести к попарным взаимодействиям.

Покажем также и пример, когда замена гиперребра на парные взаимодействия возможно: в системе рекомендаций фильмов транзакция «фильм f режиссера r вышел в прокат в году y» представляется ребром из трех вершин $e = (f, r, y)$, но гипер-ребро здесь излишне, и имеет смысл заменить его на два парных взаимодействия (f, r), (f, y).

Заметим, что в некоторых случаях порядок вершин важен: «пользователь u кликнул на рекламу r» или «реклама r была показана пользователю u»; таким образом, для представления данных используется ориентированный гиперграф.

Каждая вершина $v \in V$ имеет *модальность* $m = \mu(v)$ из заданного конечного множества модальностей M. Каждая вершина имеет ровно одну модальность таким образом множество вершин можно разбить на непересекающиеся подмножества:

$$V = \bigsqcup_{m \in M} V_m, \qquad V_m = \{v \in V \colon \mu(v) = m\}.$$

В классических тематических моделях, как было упомянуто в предыдущей главе, имеется только две модальности: документы D и слова (термины) W.

Каждое ребро $e \in E$ имеет *тип транзакции* $k = \kappa(e)$ из заданного конечного множества K, причем единственный, таким образом множество транзакций можно разбить на непересекающиеся подмножества:

$$E = \bigsqcup_{k \in K} E_k, \qquad E_k = \{e \in E \colon \kappa(e) = k\}.$$

Ребра $e = (v_1, \ldots, v_h)$ одного типа k полностью идентичны, то есть имеют одинаковую степень $h = h(k) = |e|$ и соединяют набор вершин определенной модальности, $\mu(v_j) = m_{kj}$, $j = 1, \ldots, h$.

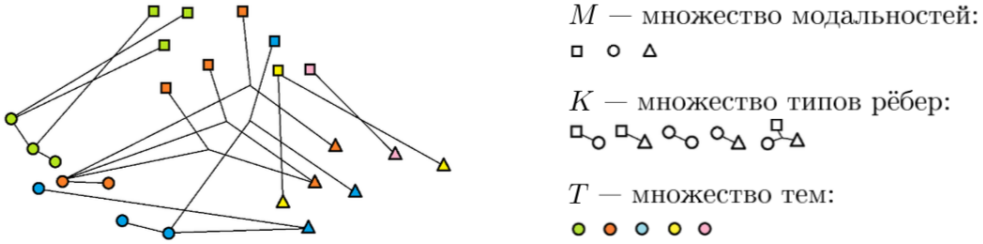

Рис 1. Пример гиперграфа с вершинам трех модальностей, ребрами-транзакциями пяти типов и пятью темами

Гиперграфовая модель работает в предположении, что транзакции $e \in E_k$ являются независимыми наблюдениями $(e,t) \in \Omega_k$, n_e – частота вхождения ребра e в выборку, и с каждой транзакцией связана своя не наблюдаемая тема $t \in T$. Причем, для каждого типа ребер k ставится в соответствие дискретное вероятностное пространство
$\Omega_k = V_{m_{k1}} \times \cdots \times V_{m_{kh}} \times T$ с функцией вероятности $p_k \colon \Omega_k \to [0,1]$.

Каждая вершина v связана с латентными темами совместной вероятностью

$$p_k(v,t) = p_k(v \,|\, t)p_k(t) = p_k(t \,|\, v)p_k(v).$$

В рамках каждой модальности нормируются вероятностные распределения:

$$\sum_{v \in V_m} p_k(v) = 1; \qquad \sum_{v \in V_m} p_k(v \,|\, t) = 1.$$

Классические тематические модели являются асимметричными: модальность документов отличается от модальности терминов, так как с документами связываются условные распределения тем $p(t \,|\, d)$, а связь терминов и тем записывается как $p(v \,|\, t)$. Будем называть модальности аналогичные модальности документов *контейнерами*. Подобная асимметрия облегчает построение модели.

Таким образом, в гиперграфовом случае, будем также использовать асимметричную модель. Предположим, что для каждого типа рёбер k первая модальность m_{k1} является контейнером (например, документом или пользователем). Обозначим через $D_k = V_{m_{k1}}$ множество всех вершин-контейнеров рёбер типа k. Для произвольной вершины-контейнера $d \in D_k$ обозначим через d_k множество таких рёбер x, что $(d,x) \in E_k$. Запись «$x \in d_k$» означает перебор всех рёбер с вершиной-контейнером d, что в классических тематических моделях соответствует перебору всех слов x в документе d.

Обозначим предположения, в рамках которых, будет задана вероятностная модель порождения транзакций:

1. Распределение тем в вершине-контейнере d не зависит от типа ребра: $p_k(t \mid d) = p(t \mid d)$ для всех $d \in D_k$.

2. Распределения вершин в темах $p_k(v \mid t)$ могут быть различны для всех типов рёбер.

3. $p_k(x \mid t) = \prod\limits_{v \in x} p_k(v \mid t)$

Пункт 1, является некоторым обобщением идеи, заложенной в мультимодальных тематических моделях, распределение тем в документе одинаково для любой модальности.

Пункт 2, говорит о том, что распределение токенов для одной и той же темы могут отличатся в зависимости от модальности. Например распределения слов в отзывах на услуги, и в рекламе тех же самых услуг являются совершенно разными. При дополнительном ограничении на похожесть данных распределений, можно ввести регурялизатор.

Под пунктом 3, вводится гипотеза условной независимости вершин в рёбрах (d, x)

Гиперграфовая тематическая модель выражает вероятности появления транзакций через распределения, связанные с их вершинами:

$$p_k(d, x) = \sum_{t \in T} p_k(x \mid t)\, p_k(t \mid d) = p_k(d) \sum_{t \in T} p(t \mid d) \prod_{v \in x} p_k(v \mid t) = p_k(d) \sum_{t \in T} \theta_{td} \prod_{v \in x} \varphi_{vtk}.$$

Параметрами этой модели являются условные вероятности вершин в темах $\varphi_{vtk} = p_k(v \mid t)$, нормированные по каждой модальности $v \in V_m$, и условные вероятности тем в контейнерах $\theta_{td} = p(t \mid d)$. Вероятности $p_k(d)$ оцениваются по наблюдаемым данным и не зависят от параметров модели:

$$p_k(d) = \frac{\sum\limits_{x \in d_k} n_{dx}}{\sum\limits_{d' \in D_k} \sum\limits_{x \in d'_k} n_{d'x}}.$$

Для того чтобы, определить гиперграфовую тематическую модель необходимо задать:

1) ориентированный гиперграф $\Gamma = \langle V, E \rangle$,

2) множество модальностей M,

3) каждой вершине поставить в соответствие модальность $\mu\colon V \to M$,

4) множество типов ребер K,

5) каждому ребру поставить в соответствие тип $\kappa\colon E \to K$,

6) множество тем T,

7) вероятностное пространство Ω_k с распределением p_k для каждого $k \in K$,

8) параметры модели $\varphi_{vtk} = p_k(v \mid t)$ и $\theta_{td} = p(t \mid d)$.

Модель PLSA является частным случаем гиперграфовой модели: имеется две модальности документы и термины, и всего один тип ребер – вхождения термина в документ.

3.1 EM-алгоритм

Для оптимизации параметров модели, по аналогии с классическими тематическими моделями, используем принцип максимума правдоподобия для каждого типа ребер k. Таким образом, для каждого типа ребер выписывается логарифм правдоподобия, а для записи их под один функционал используется их взвешенная сумму с весами τ_k, которую и будем максимизировать отбросив слагаемые вида $\tau_k n_{dx} \ln p_k(d)$:

$$L(\Phi, \Theta) = \sum_{k \in K} \tau_k \sum_{d \in D_k} \sum_{x \in d_k} n_{dx} \ln \sum_{t \in T} \theta_{td} \prod_{v \in x} \varphi_{vtk} \to \max.$$

Так же вводятся дополнительные ограничения по средствам аддитивного регуляризатора $R(\Phi, \Theta)$, что позволит сделать решение более устойчивым [1, 2].

Теперь мы имеем все для постановки задачи построения гиперграфовой тематической модели:

$$\begin{aligned}
&\sum_{k \in K} \tau_k \sum_{d \in D_k} \sum_{x \in d_k} n_{dx} \ln \sum_{t \in T} \theta_{td} \prod_{v \in x} \varphi_{vtk} + R(\Phi, \Theta) \to \max; \qquad (1)\\
&\sum_{v \in V_m} \varphi_{vtk} \in \{0, 1\}, \qquad \varphi_{vtk} \geqslant 0, \qquad k \in K,\ m \in M,\ t \in T.\\
&\sum_{t \in T} \theta_{td} \in \{0, 1\}, \qquad \theta_{td} \geqslant 0, \qquad t \in T,\ d \in D_k,\ k \in K;
\end{aligned}$$

Ограничения-равенства предусматривают возможность обнуления распределений. Если $\varphi_{vtk} = 0$ для всех $v \in V_m$, то тема t не участвует в порождении рёбер типа k с вершинами модальности m. Если $\theta_{td} = 0$ для всех $t \in T$, то считается, что тематическая модель не в состоянии описать содержимое контейнера d.

Введем оператор неотрицательного нормирования, который преобразует произвольный вектор $(a_i)_{i \in I}$ в вектор вероятностей дискретного распределения:

$$\operatorname*{norm}_{i \in I} a_i = \frac{\max\{a_i, 0\}}{\sum\limits_{j \in I} \max\{a_j, 0\}}, \text{ для всех } i \in I,$$

причем в случае, когда $a_i \leqslant 0$ для всех $i \in I$, будем полагать $\operatorname*{norm}\limits_{i \in I} a_i = 0$.

Для решения задачи (1) будем использовать регуляризованный EM алгоритм, на каждой итерации которого выполняется два шага: E и M.

На E-шаге (expectation) для каждого наблюдаемого ребра (d, x) гиперграфа вычисляется распределение тем $p_{tdx} = p(t \mid d, x)$ по формуле Байеса:

$$p_{tdx} = \operatorname*{norm}_{t \in T} \left(\theta_{td} \prod_{v \in x} \varphi_{vtk} \right). \qquad (2)$$

На M-шаге (maximization) полученные значения вспомогательных переменных p_{tdx} используются для оценивания параметров модели:

$$\varphi_{vtk} = \underset{v \in V_m}{\operatorname{norm}}\left(n_{vtk} + \varphi_{vtk}\frac{\partial R}{\partial \varphi_{vtk}}\right); \qquad n_{vtk} = \sum_{d \in D_k}\sum_{x \in d_k}[v \in x]\,\tau_k n_{dx} p_{tdx}; \tag{3}$$

$$\theta_{td} = \underset{t \in T}{\operatorname{norm}}\left(n_{td} + \theta_{td}\frac{\partial R}{\partial \theta_{td}}\right); \qquad n_{td} = \sum_{k \in K}\sum_{x \in d_k}\tau_k n_{dx} p_{tdx}; \tag{4}$$

где n_{vtk} является суммарным весом ребер типа k, содержащих вершину v и относящихся к теме t; n_{td} является суммарным весом всех ребер для вершины-контейнера d, относящихся к теме t.

Тема $t \in T$ в модальности $m \in M$ называется регулярной для типа ребер $k \in K$, если существует такая вершина $v \in V_m$, что

$$n_{vtk} + \varphi_{vtk}\frac{\partial R}{\partial \varphi_{vtk}} > 0. \tag{5}$$

Вершина-контейнер называется регулярной, если существует такая тема $t \in T$, что

$$n_{td} + \theta_{td}\frac{\partial R}{\partial \theta_{td}} > 0. \tag{6}$$

Для модели условия регулярности показывают лишь то, что воздействие регуляризатора не сильно искажают модель. Для частных случаев гиперграфовой модели, таких как PLSA и LDA, данные условия выполняются всегда.

Для тем, которые не являются регулярными, предлагается вводить $\varphi_{vtk} = 0$ для всех $v \in V_m$, где t – не регулярная тема. Таким образом, данный подход можно интерпретировать так: данная тема не участвует в порождении ребер типа k.

Для вершин-контейнеров, которые не являются регулярными, положим $\theta_{td} = 0$ для всех $t \in T$, где d не регулярная вершина-контейнер. Таким образом, вершина-контейнер d не может быть описана тематической моделью.

Покажем, что EM алгоритм (2)–(4) является методом простых итераций для решения системы уравнений эквивалентной условиям KKT для задачи (1).

Теорема 1. Если функция $R(\Phi, \Theta)$ непрерывно дифференцируема и (Φ, Θ) — точка локального максимума задачи (1), и выполняются условия регулярности (5)–(6), то выполняется система уравнений (2)–(4) относительно параметров модели φ_{vtk}, θ_{td} и вспомогательных переменных p_{tdx}, n_{td} и n_{vtk}.

Доказательство. Воспользуемся условиями KKT. Запишем лагранжиан оптимизационной задачи (1):

$$L(\Phi,\Pi)=\sum_{x\in X}\tilde{n}_x\ln\sum_{t\in T}\pi_t\prod_{v\in x}\varphi_{vt}+R(\Phi,\Pi)-$$
$$-\mu\Big(\sum_{t\in T}\pi_t-1\Big)-\sum_{t\in T}\mu_t\pi_t-\sum_{m\in M}\sum_{t\in T}\lambda_{mt}\Big(\sum_{v\in V_m}\varphi_{vt}-1\Big)-\sum_{m\in M}\sum_{t\in T}\sum_{v\in V_m}\lambda_{mvt}\varphi_{vt}$$

Продифференцируем лагранжиан и приравняем к нулю производные:

$$\frac{\partial L}{\partial\varphi_{vtk}}=\sum_{d\in D_k}\sum_{x\in d_k}[v\in x]\,\tau_k n_{dx}\frac{\theta_{td}\prod_{u\in x\setminus v}\varphi_{utk}}{p_k(d,x)}+\frac{\partial R}{\partial\varphi_{vtk}}-\lambda_{k\mu(v)t}-\lambda_{k\mu(v)vt}=0;$$
$$\frac{\partial L}{\partial\theta_{td}}=\sum_{k\in K}\sum_{x\in d_k}\tau_k n_{dx}\frac{\prod_{v\in x}\varphi_{vtk}}{p_k(d,x)}+\frac{\partial R}{\partial\theta_{td}}-\mu_d-\mu_{td}=0.$$

Домножим первое равенство на φ_{vtk}, а второе на θ_{td}:

$$\sum_{d\in D_k}\sum_{x\in d_k}[v\in x]\,\tau_k n_{dx}\underbrace{\frac{\theta_{td}\prod_{u\in x}\varphi_{utk}}{p_k(d,x)}}_{p_{tdx}=p(t\,|\,d,x)}+\varphi_{vtk}\frac{\partial R}{\partial\varphi_{vtk}}=\lambda_{k\mu(v)t}\varphi_{vtk};$$
$$\sum_{k\in K}\sum_{x\in d_k}\tau_k n_{dx}\underbrace{\frac{\theta_{td}\prod_{v\in x}\varphi_{vtk}}{p_k(d,x)}}_{p_{tdx}=p(t\,|\,d,x)}+\theta_{td}\frac{\partial R}{\partial\theta_{td}}=\mu_d\theta_{td}.$$

Перепишем условия в терминах n_{vtk} из (3) и n_{td} из (4):

$$\varphi_{vtk}\lambda_{k\mu(v)t}=n_{vtk}+\varphi_{vtk}\tfrac{\partial R}{\partial\varphi_{vtk}}. \tag{7}$$
$$\theta_{td}\mu_d=n_{td}+\theta_{td}\tfrac{\partial R}{\partial\theta_{td}}; \tag{8}$$

Случай 1. $\lambda_{kmt}\leqslant 0$, в данном случае условие регулярности (5) не выполняется, таким образом, положим, $\varphi_{vtk}=0$ для всех $v\in V_m$. Случай 2. $\lambda_{kmt}\geq 0$, то обе части равенства (7) неотрицательны. Запишем все оба случая под одну формулу:

$$\varphi_{vtk}\lambda_{k\mu(v)t}=\max\big\{n_{vtk}+\varphi_{vtk}\tfrac{\partial R}{\partial\varphi_{vtk}},0\big\}. \tag{9}$$

Действуем по аналогии: Случай 1. $\mu_d\leqslant 0$, то условие регулярности (6) не выполняется, положим, $\theta_{td}=0$ для всех $t\in T$. Случай 2. $\mu_d\geq 0$, то обе части равенства (8) неотрицательны. Запишем все оба случая под одну формулу:

$$\theta_{td}\mu_d=\max\big\{n_{td}+\theta_{td}\tfrac{\partial R}{\partial\theta_{td}},0\big\}. \tag{10}$$

После нехитрых манипуляций, и применения условий нормировки, получаем выражения для двойственных переменных:

$$\mu_d=\sum_{t\in T}\max\big\{n_{td}+\theta_{td}\tfrac{\partial R}{\partial\theta_{td}},0\big\};$$
$$\lambda_{kmt}=\sum_{v\in V_m}\max\big\{n_{vtk}+\varphi_{vtk}\tfrac{\partial R}{\partial\varphi_{vtk}},0\big\}.$$

Заменив двойственные переменные их значениями в (9) и (10), получим (3) и (4). Теорема доказана.

4 Эксперименты

4.1 Модельные данные

В данном разделе приведены эксперименты на модельных данных. Основной задачей подобных экспериментов является проверка работоспособности алгоритма, а также его сравнения алгоритма с другими методами восстановления параметров модели порождения данных, в том числе и сравнения устойчивости моделей относительно выбора числа тем.

Процедура генерации данных.

В первую очередь были сгенерированы матрицы $\Theta = p(t|d)$ и $\Phi_k = p_k(v|t)$ для всех $k \in K$. Причем каждый объект был отнесен к одному из нескольких классов, в зависимости от токенов в него входящих. Также стоит отметить, что для разных типов ребер отличаются распределения токенов по темам. Далее с помощью заданных матриц, генерируются транзакции (гипер-ребра) между объектами. В данном эксперименте были выбраны следующие параметры:

1. число тем 50
2. число классов 4
3. число объектов 1000
4. число типов ребер 6
5. число модальностей 3
6. сгенерировано около 10^7 транзакций

Постановка эксперимента.

В данном эксперименте будут сравниваться следующие алгоритмы:

1. PLSA
2. MultimodalARTM
3. TransARTM

Для сравнения алгоритмов в решении задачи классификации, будет использоваться стандартная метрика качества:

$$Accuracy = \frac{1}{N}\sum_{k=1}^{N}[y_k^{pred} == y_k^{true}].$$

Первый эксперимент – тестирование в идеальных условиях: число тем совпадает с заданным при генерации. На рисунке 4.1 мы отчетливо видим, что гиперграфовая модель (TransARTM) на порядок быстрее сходится к устойчивому решению по сравнению с двумя другими моделями.

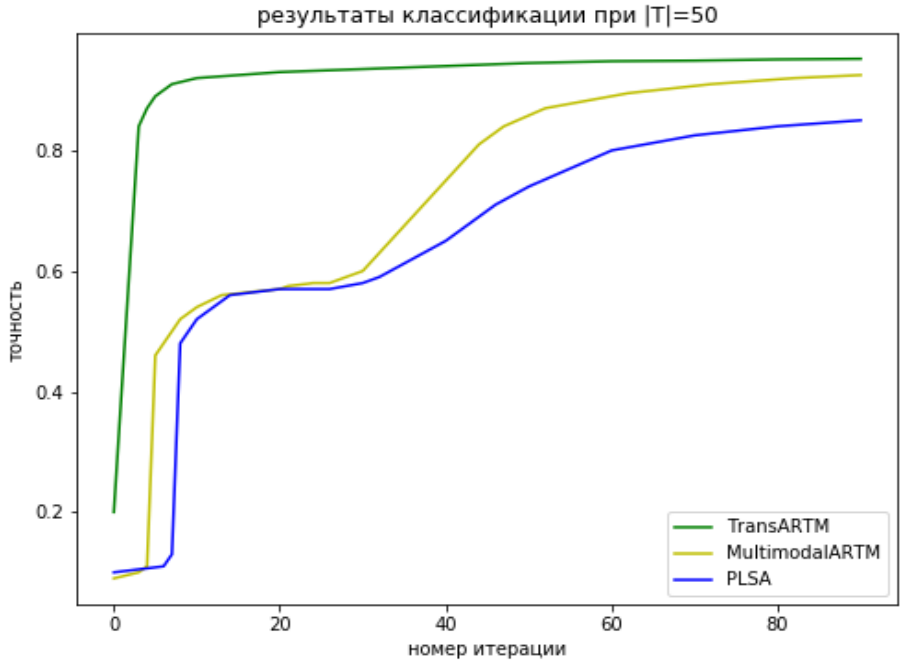

Рис. 1: |T| = 50

Попробуем приблизить эксперимент к реальным условиям – предположим, что точное число тем неизвестно, поэтому будем варьировать этот параметр от 5 до 200. Далее на рисунках 4.2-4.5 можно заметить, что TransARTM имеет явное преимущество в скорости сходимости над остальными из рассмотренных алгоритмов. Более того, решение полученное гиперграфовой моделью, устойчиво относительно числа тем, тогда как два других алгоритма имеют проблемы со сходимостью при большом количестве тем.

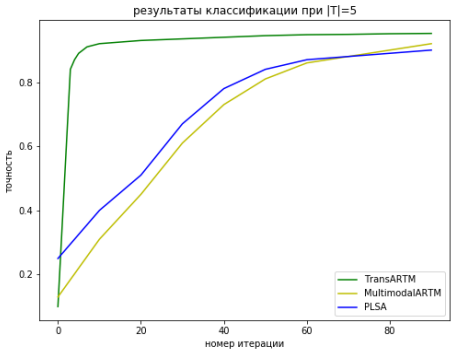

Рис. 2: |T| = 5

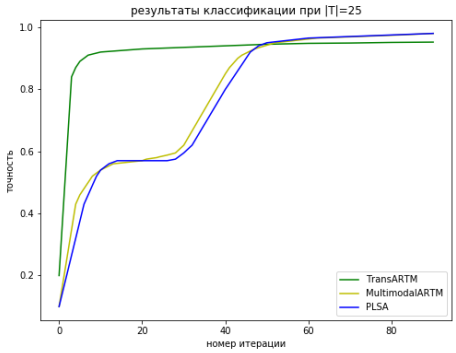

Рис. 3: |T| = 25

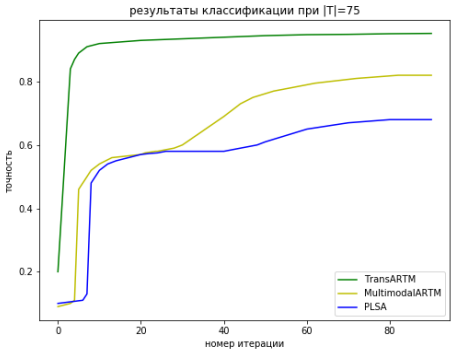

Рис. 4: |T| = 75

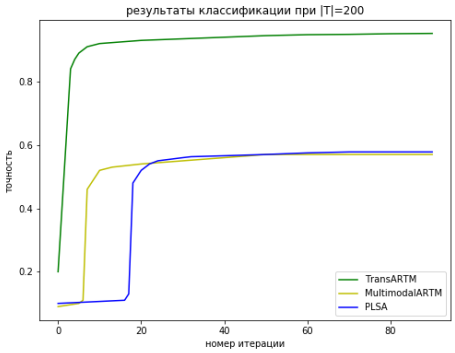

Рис. 5: |T| = 200

Вывод. На транзакционных данных TransARTM показывает наилучшие результаты как и по качеству классификации, так и по скорости сходимости среди рассмотренных моделей. Более того TransARTM оказался наиболее устойчив к выбору числа тем.

4.2 Банковские транзакции

Пусть F – конечное множество компаний, V – конечное множество видов экономической деятельности, W множество слов, встречающихся в назначениях платежей. F, V, W представляют из себя множества вершин разных типов (модальностей), таким образом множество модальностей M определено и состоит из трех элементов. Определим также два типа ребер: (f_i, v_i^{debet}, w_i) и (f_i, v_i^{credit}, w_i), компания f_i взаимодействует с компанией из отрасли v_i, имея в назначении платежа слово w_i. Два типа ребер отличаются направлением платежа.

Для данного эксперимента была сформирована выборка из транзакций порядка 50000 компаний за один календарный год (для того чтобы была учтена некоторая сезонность), из них 7932 компаний брали инвестиционные кредиты на различные цели. Итого банковских транзакций вышло порядка 90млн., при этом средняя длина предобработанного назначения

платежа – 6 токенов, т.е. ребер гиперграфа получилось около 55млн.

Было выделено 4 основных групп целей инвестиционных кредитов, они в итоге и были выбраны как целевой вектор задачи классификации: недвижимость, авто, проектная деятельность, все остальные цели были объединены в четвертый класс.

Качество решения задачи классификации проверялось лишь для клиентов, по которым известен тип инвестиционного кредита, так как выделение 5-го класса (не брал кредит) для бизнеса не имеет никакой ценности: банк и так знает, что данный клиент не является заемщиком. При этом для успешной продажи банковских продуктов, в том числе кредитов, важно знать мотивацию и потенциальные цели клиента. Данная задача классификации в банке решается по средствам градиентного бустинга над деревьями. Признаками для этого решения являются: индикаторы на основе регулярных выражений, с помощью которых, на основе экспертного мнения выделяются различные цели платежей; различные статистики по суммам платежей; дополнительные признаки о компании. Таким образом, сравнение напрямую TransARTM с текущим банковским решением, основанным на множестве различных признаков, помимо информации о транзакциях не совсем корректно, поэтому тематические профили компаний, были занесены в банковскую модель как дополнительные признаки. Новые признаки увеличили качество основного решения (см. Таблицу 1), таким образом TransARTM доказал свою применимость.

Таблица 1: Результаты решения задачи классификации

Model	**weighted f1 score**	**accuracy**	**num topics**
lightGBM	0.7331	0.6732	-
lightGBM + TransARTM	0.7327	0.6718	10
	0.7343	0.6751	50
	0.7349	**0.6759**	100
	0.7318	0.6709	200

Кроме задачи классификации, в виду ограниченности ресурсов банка, возникает задача ранжирования клиентов, по потребности в том или ином инвестиционном кредите. Для данной задачи крайне важна интерпретируемость модели, так как сотрудникам отдела продаж важно понимать, по какой причине они идут именно к этому клиенту, для того чтобы успешно продавать банковский продукты. Поэтому выбор модели ранжирования остановился на оценке удаленности от центров тематических профилей клиентов, которые брали кредиты

до этого. В качестве меры близости между тематическими профилями используется KL-дивергенция. Как меру качества ранжирования алгоритма, необходимо использовать конверсию клиентов, к сожалению на момент написания данной работы, пилотный проект еще не был запущен. В таблице 2 показаны некоторые измерения на исторических данных, но они не могут в полной мере оценить качество модели.

Таблица 2: Результаты решения задачи ранжирования

Model	**map@5000**	**map@10000**	**num topics**
top by transaction sum	0.00043	0.00041	-
TransARTM	0.0047	0.0029	50
	0.0051	0.0032	100
	0.0049	0.0032	200

5 Заключение

В экспериментах на модельных данных показано преимущество TransARTM, над PLSA и MultiModalARTM на транзакционных данных. TransARTM показал себя лучше как и по скорости сходимости модели, так и по качеству восстановления исходной матрицы Θ. TransARTM также оказался более устойчив к выбору числа тем, и даже при заведомо большом параметре $|T|$ имел быструю сходимость, в отличии от других рассмотренных алгоритмов.

На реальных данных, алгоритм позволил улучшить текущее банковское решение задачи классификации в потребности инвестиционных кредитов. Кроме того, модель показала работоспособность в качестве алгоритма ранжирования клиентов.

На защиту в данной работе выносятся следующие результаты:

1. Гиперграфовая тематическая модель
2. Эксперименты на модельных данных
3. Решение задачи по рекомендации инвестиционных кредитов для клиентов банка

Список литературы

[1] *Vorontsov K. V.* Additive regularization for topic models of text collections // Doklady Mathematics / Pleiades Publishing. — Vol. 89. — 2014. — Pp. 301–304

[2] *Воронцов К. В.* Аддитивная регуляризация тематических моделей коллекций текстовых документов // Доклады РАН. – 2014. – Т.456 №3 – С. 268-271

[3] *Adomavicius G., Tuzhilin A.* Context-aware recommeder systems // Proceedings of the 2008 ACM Conference on Recommender Systems. – RecSys '08. – New York, NY, USA: ACM, 2008. – Pp 335-336.

[4] *Blei D. M.* Probabilistic topic models // Communications of the ACM. – 2012. – Vol. 55, no. 4. – Pp. 77-84

[5] *Yin H., Cui B., Sun Y., Hu Z., Chen L.* A spatial item recommender system // ACM Transaction on Information Systems. – 2014

[6] *Воронцов К. В., Айсина Р. М.* Тематическое моделирование финансовых потоков корпоративных клиентов банка по транзакционным данным. 2017.

[7] *Hofmann, T.* Unsupervised Learning by Probabilistic Latent Semantic Analysis // Machine Learning, 42, 177–196, 2001 Kluwer Academic Publishers.

[8] *David M. Blei, Andrew Y. Ng, Michael I. Jordan* Latent Dirichlet Allocation // Journal of Machine Learning Research 3 (2003) 993-1022