МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ РОССИЙСКОЙ ФЕДЕРАЦИИ

МОСКОВСКИЙ ФИЗИКО ТЕХНИЧЕСКИЙ ИНСТИТУТ (ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ)

ФАКУЛЬТЕТ ИННОВАЦИЙ И ВЫСОКИХ ТЕХНОЛОГИЙ

КАФЕДРА АНАЛИЗА ДАННЫХ

Выпускная квалификационная работа по направлению
010400 «Прикладная математика и информатика»

Использование графовой структуры в тематическом моделировании

Студент 6 курса Булатов В. Г.

Научный руководитель Бунина Е.И.

Москва 2016

Содержание

Аннотация

В данной работе было предложено несколько моделей, в рамках вероятностного тематического моделирования учитывающих текстовую информацию и информацию о связях между документами.

Предложенные модели сформулированы на языке аддитивной регуляризации тематических моделей, поэтому они гибки и допускают комбинирование друг с другом или с другими моделями.

Предложенные модели тестируются на корпусе из названий научных статей с ссылками по признаку соавторства.

Автор благодарит Воронцова Константина Вячеславовича за консультации и обсуждения.

1 Введение

В современном мире мы наблюдаем множество текстовой информации, каким-то образом организованной в виде сети. Примерами такой организации являются научные статьи, связанные друг с другом через сеть цитирований и соавторств; тексты в блогах, связанные друг с другом через репосты и ссылки; сообщения в Твиттере, которые могут быть ответами или ретвитами.

Задача анализа текстовой информации в настоящее время достаточно хорошо изучена. Одним из наиболее популярных подходов к ней является *вероятностное тематическое моделирование*, использующее понятие *темы* для описания процесса генерации документов. Тематическое моделирование не позволяет напрямую учесть граф связей между документами, но существует ряд работ, пытающихся обойти это ограничение.

Как правило, такие работы вводят более сложную модель генерации данных, призванную учесть какие-то требования, накладываемые на решение. К сожалению, такие модели затруднительно сочетать как и друг с другом, так и с моделями, не затрагивающими связи между документами (но каким-то образом улучшающими качество анализа текстовой информации). Кроме того, предложенные решения плохо распараллеливаются и вследствие чего плохо масштабируются.

В настоящей работе предложен и формально обоснован гибкий способ учёта графовой информации, позволяющий комбинировать друг с другом широкий класс подходов, а также совместимый с онлайновым пакетным EM-алгоритмом.

Основной фокус работы — переход от задачи максимизации правдоподобия к задаче максимизации регуляризованного правдоподобия. Регуляризацию модели можно интерпретировать как учёт какой-то дополнительной информации о природе коллекции документов и о критериях качества, по которым будет оцениваться решение.

Как будет показано в дальнейшем, регуляризационный подход позволяет позволяет учитывать информацию о взаимосвязях документов гибким, естественным и вычислительно эффективным способом.

Работа организована следующим образом.

В третьем разделе описывается модель PLSA и два её обобщения: многомодальная тематическая модель и подход APTM. Вводится ряд нужных обозначений и фактов.

Четвертый раздел посвящен различным способам учёта графовой структуры. Рассматриваются подходы из литературы, предлагаются и анализируются альтернативные подходы.

Пятый раздел содержит описание эксперимента и его результаты.

В шестом разделе приводятся и интерпретируются некоторые факты общего характера.

2 Определения и обозначения

2.1 Постановка задачи PLSA

Имеется W — конечное множество слов (называемое *словарём*), D — конечное множество документов (называемое *коллекцией*). Для каждого $d \in D$ и $w \in W$ определим целочисленную величину n_{dw}, обозначающую число вхождений слова w в документ d.

В рамках задачи *тематического моделирования* считается, что помимо наблюдаемых данных (n_{dw}) имеются ещё и конечное множество переменных T (называемых *темами*), и на пространстве $W \times D \times T$ задано вероятностное распределение $p(w, d, t)$.

Процесс генерации документа выглядит так:

1. Из распределения $p(t \mid d)$, обозначаемого θ_{td} выбирается тема t
2. Из распределения $p(w \mid t)$, обозначаемого ϕ_{wt} выбирается слово w

Используя введённые обозначения, можно написать, что $p(w \mid d) = \sum_t \phi_{wt}\theta_{td}$. Задача классического тематического моделирования состоит в нахождении этих распределений Θ и Φ.

Критерием качества тематической модели является логарифм правдоподобия:

$$L = \sum_d \sum_w n_{dw} \log \sum_{k=1}^{K} \theta_{td}\phi_{wt} \to \max_{\Theta,\Phi}$$

2.2 EM-алгоритм для PLSA

Можно видеть, что в задаче PLSA имеются наблюдаемые данные (слова внутри документа, n_{dw}) и скрытые параметры (к какой теме относится слово). Скрытую информацию удобно обозначить через через совокупность индикаторных переменных z_{tdw}, где z_{tdw} равна 1 в случае, когда слово w в документе d принадлежит теме t, и 0 иначе.

Как правило, в подобных задачах максимизация правдоподобия происходит при помощи EM-алгоритма. EM-алгоритм оперирует не только

с параметрами распределения Φ и Θ (удобно для краткоости обозначить эти параметры за ψ), но и со скрытыми переменными, стремясь максимизировать уже не L, а математическое ожидание L по скрытым переменным.

Если обозначить «истинную» тему слова w_i через t_i^*, то нетрудно записать формулу для L^c, правдоподобия полных данных. В дальнейшем будет удобно переписать её в виде, использующем скрытые параметры z_{tdw}.

$$L^c = \log \Pr(N \mid Z, \Theta) =$$

$$\sum_d \sum_{w_i} \log p(d, w_i, t_i^*) = \sum_d \sum_{w_i} \sum_{t=t_i^*} \log p(d, w_i, t) =$$

$$\sum_d \sum_{w_i} \sum_t z_{td,w_i} \log p(d, w_i, t) =$$

$$\sum_d \sum_{w_i} \sum_t z_{td,w_i} \log[p(w_i, t \mid d) p(d)] =$$

$$\sum_d \sum_{w_i} \sum_t z_{td,w_i} (\log[\phi_{wt}\theta_{td}] + \log p(d)) =$$

$$\sum_d \sum_{w_i} \sum_t z_{td,w_i} \log[\phi_{wt}\theta_{td}] + \sum_d \sum_{w_i} \sum_t z_{td,w_i} \log p(d) =$$

$$\sum_d \sum_{w_i} \sum_t z_{td,w_i} \log[\phi_{wt}\theta_{td}] + \text{const}$$

EM-алгоритм итеративно приближает свои оценки ψ к локальному максимуму правдоподобия, чередуя E-шаги и M-шаги.

На E-шаге (estimation) вычисляется целевая функция Q, равная математическому ожиданию L по Z. Используются текущие оценки ψ (т.е. для Φ и Θ).

$$Q(\psi \mid \psi^{(t)})_{\text{PLSA}} = \mathbb{E}\big[L^c\big] =$$

$$\mathbb{E}\big[\sum_d \sum_{w_i} \sum_t z_{td,w_i} \log[\phi_{wt}\theta_{td}] + \text{const}\big] \propto$$

$$\sum_d \sum_{w_i} \sum_t p_{td,w_i} \log[\phi_{wt}\theta_{td}] =$$

$$\sum_d \sum_w n_{dw} \sum_t p_{tdw} \log[\phi_{wt}\theta_{td}]$$

Видно, что в случае PLSA Q целиком задаётся при помощи вспомогательных переменных p_{tdw} — матожиданий значений скрытых переменных z_{tdw}. Поэтому на практике E-шаг заключается просто в вычислении значений p_{tdw} по формуле Байеса. Запишем эту формулу, используя оператор нормировки $\underset{t\in T}{\text{norm}}$:

$$p_{tdw} = \underset{t\in T}{\text{norm}}(\phi_{wt}\theta_{td})$$

На M-шаге (maximisation) находится значение ψ, максимизирующее значение Q. Решение этой задачи можно выписать явно:

$$\theta_{td} = \underset{w\in W}{\text{norm}}(\sum_w n_{dw}p_{tdw}) = \underset{w\in W}{\text{norm}}(n_{td})$$

$$\phi_{wt} = \underset{d\in D}{\text{norm}}(\sum_d n_{dw}p_{tdw}) = \underset{d\in D}{\text{norm}}(n_{wt})$$

Величину n_{td} можно проинтерпретировать, как математическое ожидание числа слов по теме t внутри документа d. Величина n_{wt} аналогично интерпретируется как ожидаемое число раз, когда термин w был употреблён в связи с темой t.

В случае различных модификаций PLSA общих формул для обновления значений θ может не существовать. В этом случае используется GEM-алгоритм, где вместо максимизации происходит просто улучшение Q какими-либо численными методами (и поэтому может быть важным умение вычислять Q).

2.3 Регуляризация

Можно рассмотреть обобщение задачи PLSA, где помимо правдоподобия имеется ещё I критериев R_i (называемых *регуляризаторами*), и требуется максимизировать не правдободобие, а линейную комбинацию правдоподобия и регуляризаторов.

$$L + \tau R = \sum_d \sum_w n_{dw} \log \sum_{k=1}^{K} \theta_{td}\phi_{wt} + \sum_i \tau_i R_i(\Theta, \Phi) \to \max_{\Theta,\Phi}, \qquad (1)$$

где $\{\tau_i\}_i^I$ — множество неотрицательных *коэффициентов регуляризации.*

При определённых условиях возможно решение обобщённой задачи (1) посредством итерационного метода, структурно похожего на EM-алгоритм, и позволяющего единообразно учитывать различные регуляризаторы.

Теорема 1. *Назовём тему t регулярной, если найдётся термин $w \in W$, для которого справедливо $n_{wt} + \phi_{wt}\frac{\partial R}{\partial \phi_{wt}} > 0$. Назовём документ d регулярным, если найдётся тема $t \in T$, для которой справедливо $n_{dt} + \theta_{td}\frac{\partial R}{\partial \theta_{td}} > 0$.*

Если функция $R(\Phi, \Theta)$ непрерывно дифференцируема и (Φ, Θ) — точка локального экстремума задачи (1), то для всех регулярных тем t и регулярных документов d справедлива система уравнений:

$$p_{tdw} = p(t \mid d, w) = \underset{t \in T}{\text{norm}} \left(\theta_{td}\phi_{wt}\right) \qquad (2)$$

$$\theta_{td} = \underset{t \in T}{\text{norm}} \left(n_{dt} + \theta_{td}\frac{\partial R}{\partial \theta_{td}}\right) \quad (3) \quad \phi_{wt} = \underset{w \in W}{\text{norm}} \left(n_{wt} + \phi_{wt}\frac{\partial R}{\partial \phi_{wt}}\right) \quad (4)$$

Эта теорема доказывается непосредственным применением условий Каруша-Куна-Таккера. Подробно её доказательство изложено в [16].

Главное следствие этой теоремы состоит в том, что задачу (1) возможно решить при помощи итеративного процесса, поочерёдно применяя формулы (2), (4), (3). Полученная процедура интерпретируется, как EM-алгоритм.

Заметим, что если рассмотреть линейную комбинацию регуляризаторов $R = \sum_i \tau_i R_i(\Phi, \Theta)$, то EM-алгоритм, оптимизирующий искомое регуляризированное правдоподобие можно построить простой модификацией M-шага:

$$\theta_{td} = \underset{t}{\text{norm}}\left(n_{dt} + \sum_i \tau_i \theta_{td} \frac{\partial R_i}{\partial \theta_{td}}\right) = \underset{t}{\text{norm}}\left(n_{dt} + \sum_i \tau_i r_{td}^{(i)}\right)$$

$$\phi_{wt} = \underset{t}{\text{norm}}\left(n_{wt} + \sum_i \tau_i \phi_{wt} \frac{\partial R_i}{\partial \phi_{wt}}\right) = \underset{t}{\text{norm}}\left(n_{wt} + \sum_i \tau_i r_{wt}^{(i)}\right)$$

Отсюда возникает подход *аддитивной регуляризации тематических моделей* (ARTM), решающий задачу многокритериальной оптимизации. Основной принцип ARTM состоит в рассмотрении линейной комбинации регуляризаторов, а затем построении M-шага как линейной комбинации *регулязационных добавок* r_{wt} или r_{td}. Этот подход гибок, единообразен и достаточно вычислительно эффективен.

Одна из целей настоящей работы состоит в разработке способа учёта графовой информации, совместимого с подходом ARTM.

2.4 Многомодальное тематическое моделирование

Существует дальнейшее обобщение задачи тематического моделирования.

В классическом тематическом моделировании считается, что документы содержат лишь слова $w \in W$, однако на практике о документе часто известна какая-то дополнительная информация, представимая в виде меток из некого конечного множества другой природы.

Примерами таких меток могут быть авторы документа, категории, дата и места публикации, реклама на странице.

В рамках подхода тематического моделирования можно учесть информацию о метках, если в качестве вероятного пространства рассмотреть $[W^1 \times \ldots W^m] \times T \times D$ (вместо $D \times T \times W$), где W^k — это конечные множества, называемые *модальностями*.

Вероятность появления термина w из k-й модальности в документе d задаётся следующей формулой: $p(w^k \mid d) = \sum_t \phi_{wt}^k \theta_{td}$, а общее правдоподобие представляется так:

$$L(\Phi^m, \Theta) = \sum_m \tau_m \sum_{d \in D} \sum_{w \in W^m} n_{dw} \ln p(w \mid d) \to \max,$$

где для общности введены коэффициенты τ_m, показывающие "вес" модальности m.

EM-алгоритм для этой задачи структурно похож на EM-алгоритм для PLSA: вводятся p_{tdw}, ϕ_{wt} и θ_{td}, их оценки итеративно обновляются. На Φ накладываются следующие ограничения: $\forall k \forall t \sum_{w^k} \phi_{w^k,t} = 1$.

3 Учёт графической структуры

Пусть теперь имеется взвешенный (ориентированный либо неориентированный) граф $G = \langle D, E \rangle$, несущий информацию о связях между документами. Обозначим за $w(u, v)$ вес ребра между вершинами u и v.

Иногда будет удобно обращаться с графом, как с локальной информацией внутри документа. В таких ситуациях мы будем использовать следующие обозначения: $[d]$ — метка "текущий документ ссылается на документ d"; $n_{[d]}$ — число ссылок на документ d (входящая степень вершины графа); n^{links} — суммарное число всех ссылок; n_t^{links} — ожидаемое число ссылок по теме t, $n_{[d]t}$ — ожидаемое число ссылок на d, связанных с темой t.

3.1 Известные тета-регуляризаторы

Здесь мы рассмотрим работы, в которых вводятся тета-регуляризаторы.

Как правило, принцип их действия сводится к сглаживанию тематических распределений соседей.

Для всех из них M-шаг для Φ происходит по формулам 4 без изменений, а целевая функция Q выглядит следующим образом:

$$Q(\psi \mid \psi^{(t)}) = Q(\psi \mid \psi^{(t)})_{\text{PLSA}} + R(\Theta, G) \to \max$$

3.1.1 netPLSA

Введён в [8]. Роль регуляризатора выполняет взвешенная квадратичная невязка распределений соседних тем:

$$R = \frac{1}{2} \sum_{(u,v) \in E} w(u, v) \sum_{t=1}^{T} (\theta_{tu} - \theta_{tv})^2$$

M-шаг для Φ происходит без изменений; M-шаг для Θ отдельно оптимизирует правдоподобие и квадратичную невязку.

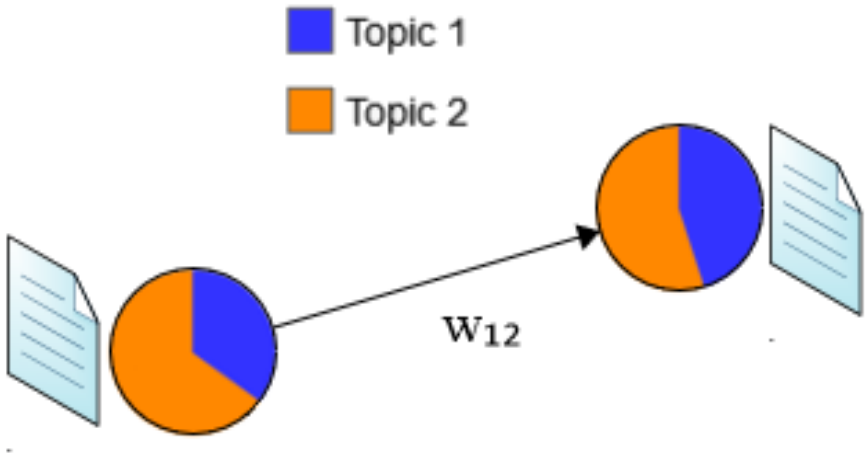


Рис. 1: Общая идея: документы D_1 и D_2 имеют схожее тематическое распределение.

Интуитивный смысл:
"Вряд ли статья, в которой нет ни одного архитектурного термина, будет цитировать статью, целиком посвящённую архитектуре"
Чем больше вес ребра между документами, тем сильнее становится требование о схожести тематических рапспределений.

Листинг 1: M-шаг для netPLSA

```
Q0 = Q(Φ, Θ)
for t, d in topics, documents:
    θ_td = norm_t n_dt
for iter in range(max_iters):
    Q[iter] = Q(Φ, Θ)
    if Q[iter] > Q0:
        # Q улучшилось по сравнению с началом цикла,
        # заканчиваем итерации
        break;
    if Q[iter] < Q[iter−1]:
        # Q начало ухудшаться,
        # заканчиваем итерации
        break;
    for t, d in topics, documents:
        r_td = norm_t ( Σ_{(d,v)∈E} w(d, v)θ_tv )
        θ_td = norm_t(θ_td + τ r_td)
```

Сначала Θ пересчитывается посредством известной формулы $\theta_{td} \propto n_{td}$, что увеличивает нерегуляризованное правдоподобие, но может ухудшить квадратичную невязку. Затем некоторое число раз происходит преобразование, оптимизирующее значение регуляризатора.

Эту процедуру можно проинтерпретировать, как градиентный спуск (хотя скорее там просто проводится линия, соединяющая точки оптимума R и Q_{PLSA}, а затем выбирается лежащая на этой линии точка).

Более подробно алгоритм приведён в листинге 1.

Можно рассмотреть менее вычислительно затратную модификацию netPLSA, ограничив число внутренних итераций M-шага единицей. Проведя ряд экспериментов, я пришёл к выводу, что при правильном подборе коэффицента регуляризации модификация с одним шагом показывают качество, сравнимое с качеством изначальной модификации (но при этом

намного более вычислительно эффективна).

3.1.2 iTopic

Введён в [13]. Оптимизируется взвешенная кросс-энтропия между тематическими распределениями соседних документов (это позволяет учитывать несимметричные связи):

$$R = \sum_{(u,v)\in E} w(u,v) \sum_{t=1}^{T} \theta_{vt} \log \theta_{ut}$$

Формулы для M-шага получены применением условий Каруша-Куна-Такера (аналогично (1)). M-шаг для Θ выглядит следующим образом:

$$\theta_{td} = \underset{t}{\operatorname{norm}} \left(n_{dt} + \tau \sum_{(d,v)\in E} w(d,v)\theta_{tv} + \tau\theta_{td} \sum_{(v,d)\in E} w(v,d) \log \theta_{tv} \right)$$

По какой-то причине авторы отказались от слагаемого с логарифмом, и поэтому итоговая формула выглядит так:

$$\theta_{td} = \underset{t}{\operatorname{norm}} \left(n_{dt} + \tau \sum_{(d,v)\in E} w(d,v)\theta_{tv} \right)$$

3.1.3 Originator or Propagator?

Введено в [17] в контексте задачи анализа поведения пользователей Twitter.

Считается, что в каждом взаимодействии пользователь играет одну из двух социальный ролей: создателя контента или распространителя чужого контента. Поэтому пользователю u соответствуют два тематических распределения: вероятность возникновения темы t в оригинальном сообщении $\theta_{tu} = p(t \mid u)$ и вероятность темы t в ретвите $\theta'_{tu} = p'(t \mid u)$.

Далее пусть n_a — число оригинальных сообщений пользователя a; r_a — число всех его ретвитов, а r_{ab} — сколько раз b сделал ретвит сообщения a. Вводятся три меры близости пользователей:

$$\mathrm{sim}_1(a,b) = \frac{r_{ab}}{n_a + r_b - r_{ab}}$$

$$\text{sim}_2(a,b) = \frac{\sum_c r_{ac} r_{bc}}{\sqrt{\sum_c r_{ac}^2}\sqrt{\sum_c r_{bc}^2}}$$

$$\text{sim}_3(a,b) = \frac{\sum_c r_{ca} r_{cb}}{\sqrt{\sum_c r_{ca}^2}\sqrt{\sum_c r_{cb}^2}}$$

Два последних выражения представляют собой просто Cosine Similarity между векторами источников /распространителей контента.

Используя их в качестве весов, определяются три регуляризатора:

- Если Боб сделал много ретвитов Алисы, то их тематики должны быть схожими:
  $R_1 = \sum_{a,b} \text{sim}_1(a,b) \sum_t (\theta'_{ta} - \theta_{tb})^2 \to \min$

- Если Алиса и Боб часто ретвитят одних и тех же пользователей, то их тематики должны быть схожими:
  $R_2 = \sum_{a,b} \text{sim}_2(a,b) \sum_t (\theta'_{ta} - \theta'_{tb})^2 \to \min$

- Если одни и те же люди распространяют и контент Алисы, и контент Боба, то их тематики должны быть схожими:
  $R_3 = \sum_{a,b} \text{sim}_3(a,b) \sum_t (\theta_{ta} - \theta_{tb})^2 \to \min$

4 Связи как модальности

Все рассмотренные выше регуляризаторы каким-то образом модифицируют распределение θ_{td} документа d, используя для этого информацию о тематических распределениях в его соседях v_i. Это приводит к необходимости держать в памяти матрицу Θ целиком. Это приводит к трудностям, когда требуется проанализировать большую коллекцию документов.

В частности, Θ-регуляризаторы невозможно напрямую применять в рамках пакетного онлайнового EM-алгоритма. Этот алгоритм разбивает коллекцию документов на пакеты, обрабатываемые параллельно и связанные только посредством матрицы Φ.

Поэтому необходим альтернативный способ учёта графовой структуры между документами, который и предлагается в этой работе.

Добавим к модальности слов модальность ссылок $W^2 = D$, расширив таким образом наше вероятностное пространство до $W \times W^2 \times D \times T$.

Рассмотрим документ d и множество $adj_d = \{v \mid (d, v) \in E\}$. Для каждого $v \in adj_d$ добавим в d метку c_v с весом $w(d, v)$. Эта метка будет означать, что в документе d имеется ссылка на документ v.

Этот подход достаточно естественен. Например, в тексте большинства научных статей встречаются как слова (например, `импульс` или `взвесь`), так и ссылки (например, `[18]` или `[Иванов 2001]`). Таким образом, рёбра графа связей можно представить как набор локальных меток, а не множество глобальных свойств[1].

Обозначим за $[v]$ метку "текущий документ ссылается на документ v", чтобы подчеркнуть эту идею. Ясно, что $w(d, v) = n_{d[v]}$ (число вхождений ссылки на d в документ u).

Таким образом, каждый документ содержит в себе метки слов и метки ссылок. Общее правдоподобие коллекции записывается таким образом:

[1] Чуть более ясно это будет звучать, если использовать англоязычную терминологию: мы перешли от ссылок в графе (links) к ссылкам внутри документа (references)

$$\log L = \sum_d \sum_w n_{dw} \log \sum_{t=1}^{T} \theta_{td}\phi_{wt} + \gamma \sum_d \sum_c w(d,c) \log \sum_{t=1}^{T} \theta_{td}\phi_{ct}$$

Легко показать, что в этом случае M-шаг будет выглядеть так:

$$\theta_{td} = \underset{t}{\text{norm}} \left(n_{dt} + \gamma \sum_{(d,v)\in E} n_{d,v} p_{tdw} \right) =$$

$$\underset{t}{\text{norm}} \left(n_{dt} + \tau \sum_{(d,v)\in E} w(d,v) \frac{\theta_{td}\phi_{vt}}{\sum_k \theta_{kd}\phi_{ck}} \right)$$

А "ссылочная" Φ в этом случае будет пересчитываться следующим образом:

$$\phi_{ct} \propto n_{ct} = \sum_d w(d,c) p_{tdc} = \sum_d w(d,c) \frac{\theta_{td}\phi_{ct}}{\sum_k \theta_{kd}\phi_{ck}}$$

5 Регуляризация

Если воздействие на матрицу Θ означает сглаживание тематик соседних документов, то воздействие на матрицу Φ можно проинтерпретировать, как влияние на информацию, которую несёт наличие термина в документе.

Это соображение позволяет сформулировать аналоги описанных выше Θ-регуляризаторов, воздействующих на матрицу Φ. Рассмотрим эту методику на примере кросс-энтропийного регуляризатора $R(\Theta, G) = R = \sum_{(u,v)\in E} w(u,v) \sum_{t=1}^{T} \theta_{vt} \log \theta_{ut}$.

5.1 Распределение "тема с учётом ссылки"

Вспомним, что величина $\theta_{td} = p(t \mid d)$ означает вероятность встретить тему t, находясь внутри документа d. Можно рассмотреть похожее распределение $p(t \mid [d])$, означающее вероятность встретить тему t, находясь внутри какого-то документа, ссылающегосся на документ d.

Заметим, что

$$p(t \mid [d]) = \frac{p(t)}{p([d])} p([d] \mid t) = \frac{p(t)}{p([d])} \phi_{[d]t}$$

Далее воспользуемся частотными оценками для вероятностей: $p(t) = n_t^{\text{links}}/n^{\text{links}}$, $p([v]) = n_{[v]}/n^{\text{links}}$:

$$p(t \mid [d]) = \frac{n_t^{\text{links}}}{n_{[d]}} \phi_{[d]t}$$

Поскольку $n_{[d]}$ есть константа, равная числу всех ссылок на d (иными словами, входящей степени вершины d), то видно, что величина $p(t \mid [d])$ зависит лишь от Φ и n_t^{links}. Это позволяет заменить $R(\Theta)$ на $R(\Phi, n_t^{\text{links}})$, подставив $p(t \mid [v])$ вместо $p(t \mid v)$, а затем переписав его в терминах n_t^{links} и $\phi_{[v]t}$.

Строго говоря, Φ и n_t^{links} не являются независимыми. Это не позволяет напрямую применить теорему 1.

Существует ряд эвристических подходов к оптимизации этого функционала.

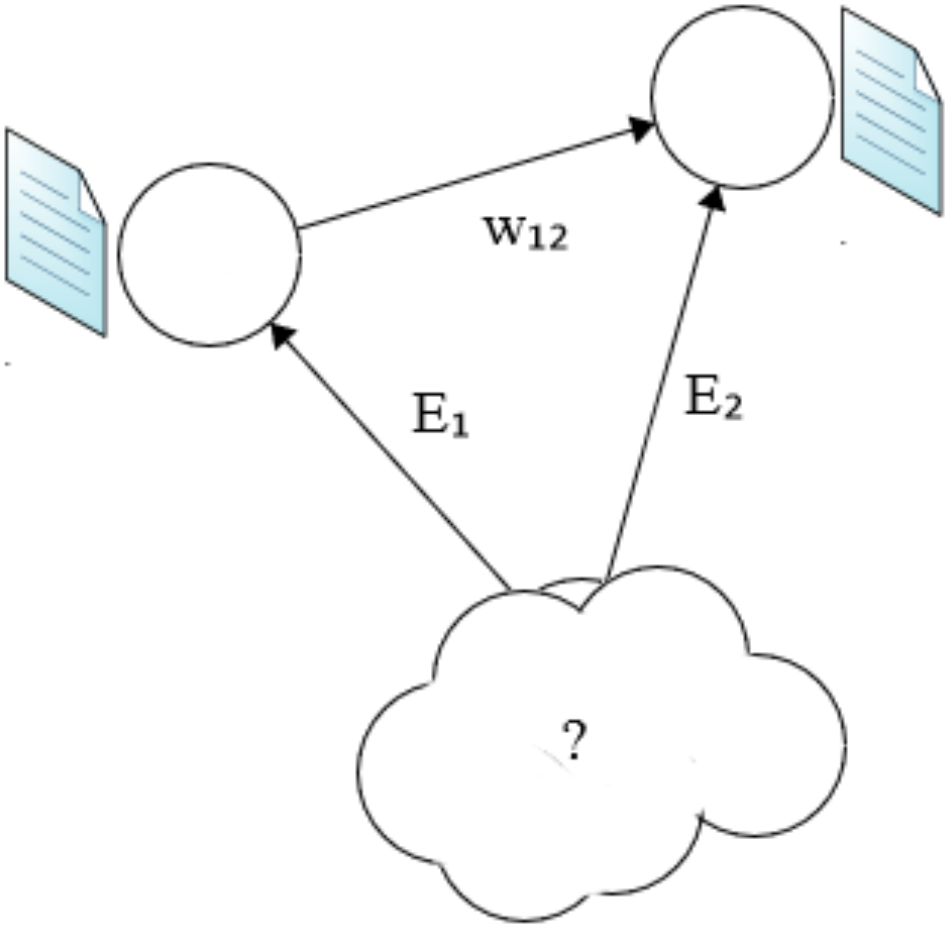


Рис. 2: Общая идея: наличие рёбер E_1 и E_2 несёт схожую информацию (т.е. позволяет сделать схожие выводы о темах внутри "нового" документа, о котором неизвестно ничего, кроме его соседей).

Интуитивный смысл:
"Если Боб подписан на Алису, то подписка Оскара на Алису и подписка Оскара на Боба должны нести схожую информацию об интересах Оскара"
Это рассуждение можно сделать более строгим, используя отношения правдоподобия.

- Модифицировать формулы M-шага для Θ-регуляризатора, заменив в них θ_{td} на $\frac{n_t \phi_{[d]t}}{n_{[d]}}$, а затем выразить из полученного соотношения $\phi_{[d]t}$

- Примененить теорему 1, некорректно считая n_t не зависящей от Φ константой

- Отдельно получить решение оптимизационной задачи для R и для PLSA. В качестве ответа выдать точку, являющуюся их взвешеной комбинацией.

Важно отметить, что все они показали худшее качество, чем обоснованная ниже методика.

5.2 Вывод формулы для градиентного спуска

В этом разделе будут обоснованы общие формулы M-шага для случая, когда регуляризатор R зависит не от ϕ_{wt}, а от ненормированных величин n_{wt}.

Лемма. *Пусть* $n_{wt} = \sum_d n_{dw} p_{tdw}$, $n_t = \sum_w n_{wt}$ *и* $Q = Q_{\text{PLSA}} + R(n_{wt})$, *где функция* $R(n_{wt})$ *непрерывно дифференцируема.*

Тогда для достаточно малых $\tau > 0$ *выполняется:*

$$Q\big(\frac{n_{wt}}{n_t} + \tau \frac{\partial R}{\partial n_{wt}}\big) > Q\big(\frac{n_{wt}}{n_t}\big)$$

Доказательство. Перепишем правдоподобие в терминах n_{wt} и перейдём от оптимизации по ϕ_{wt} к оптимизации по n_{wt}.

Введём $L'(n_{wt}, \theta_{td}) = L(\frac{n_{wt}}{\sum_w n_{wt}}, \theta_{td})$ и запишем для него функционал Q:

$$Q = \sum_{d,w} n_{dw} \sum_t p_{tdw} \log \frac{n_{wt}\theta_{td}}{\sum_w n_{wt}} + R(n_{wt}) = Q'_{\text{PLSA}} + R(n_{wt})$$

Заметим, что максимум Q'_{PLSA} по-прежнему достигается в точке $s_{wt} = \sum_d n_{dw} p_{tdw}$. Следовательно, в точке s_{wt} градиент Q'_{PLSA} зануляется. Тогда для градиента Q' справедливо:

$$\operatorname{grad} Q' = \operatorname{grad}(Q'_{\text{PLSA}} + R) = \operatorname{grad} Q'_{\text{PLSA}} + \operatorname{grad} R = \operatorname{grad} R$$

Таким образом, если сделать достаточно маленький шаг из точки n_{wt} по направлению $\operatorname{grad} R$, то значение Q' улучшится. □

Используя эту лемму, можно построить GEM-алгоритм для регуляризованного правдоподобия.

Утверждение 1. *В условиях предыдущей леммы построенный по следующим формулам итерационный процесс будет являться GEM-алгоритмом:*

$$p_{tdw} = \underset{t \in T}{\operatorname{norm}} \left(\theta_{td} \frac{n_{wt}}{\sum_w n_{wt}} \right) \tag{5}$$

$$\theta_{td} = \underset{t \in T}{\operatorname{norm}} \left(n_{dt} \right) \qquad (6) \qquad \phi_{wt} = \underset{w \in W}{\operatorname{norm}} \left(n_{wt} + \tau \frac{\partial R}{\partial n_{wt}} \right) \quad (7)$$

Доказательство. Из предыдущей леммы имеем, что если модифицировать EM-алгоритм, переписав правдоподобие в терминах n_{wt}, то описанные M-шаги будут улучшать функционал Q (что и требуется в GEM-алгоритме).

Осталось заметить, что значения скрытых переменных вычисляются по формуле $p_{tdw} = \underset{t}{\operatorname{norm}} \frac{n_{wt}\theta_{td}}{\sum_w n_{wt}}$ и завершив таким образом замену скрытых переменных. □

Отметим, что эта идея хорошо вписывается в подход APTM: формула M-шага будет выглядеть как $\phi_{wt} = \underset{w}{\operatorname{norm}}(n_{wt} + r_{wt})$ для $r_{wt} = \frac{\partial R}{\partial n_{wt}}$, а никаких существенных изменений внутри E-шага не произошло вовсе (поскольку в рамках задачи PLSA ϕ_{wt} вычисляется как $\frac{n_{wt}}{\sum_w n_{wt}}$)

Воспользуемся этой техникой для того, чтобы получить формулы M-шага для кросс-энтропийного регуляризатора. Ясно, что вычислять частные производные имеет смысл вычислять лишь для $n_{[d]t}$ (для n_{wt} они будут равны нулю):

$$\frac{\partial R}{\partial n_{[d]t}} =$$

$$\frac{\partial}{\partial n_{[d]t}} \sum_{d,v} w(d,v) \frac{n_{[v]t}}{n_{[v]}} \log\left(\frac{n_{[d]t}}{n_{[d]}}\right) +$$
$$\frac{\partial}{\partial n_{[d]t}} \sum_{u,d} w(u,d) \frac{n_{[d]t}}{n_{[d]}} \log\left(\frac{n_{[u]t}}{n_{[u]}}\right)$$

=

$$\sum_{d,v} w(d,v) \frac{n_{[v]t}}{n_{[v]}} \frac{1}{n_{[d]t}} + \sum_{u,d} w(u,d) \frac{1}{n_{[d]}} \log\left(\frac{n_{[u]t}}{n_{[u]}}\right)$$

Откуда получаем:

$$\phi_{[d]t} = \underset{[d]}{\text{norm}} \left(n_{[d]t} + \sum_{d,v} w(d,v) \frac{p(t \mid [v])}{n_{[d]t}} + \frac{1}{n_{[d]}} \sum_{u,d} w(u,d) \log(p(t \mid [u])) \right)$$

5.3 Распределение "ссылка с учётом темы"

Существует способ, более простой, чем описанный выше: просто заменить θ_{td} на $\phi_{[d]t}$ внутри регуляризатора. Несмотря на свою кажущуюся наивность, этот подход можно обосновать, как приближение байесового вывода (что будет изложено в 7.1). Иными словами,

$$R(\Phi) = \sum_{(u,v) \in E} w(u,v) \sum_{t=1}^{T} \phi_{[v]t} \log \phi_{[u]t}$$

Для этого случая формулы выводятся при помощи прямого применения теоремы 1:

$$\phi_{[d]t} = \underset{[d]}{\text{norm}} \left(n_{[d]t} + \sum_{v \in D} w(d,v) \phi_{[v]t} + \phi_{[d]t} \sum_{(v,d) \in E} w(v,d) \log \phi_{[v]t} \right)$$

5.4 Гибридный регуляризатор

Стоит обратить внимание на принципиальное отличие рассмотренных Ф-регуляризаторов от изначальных Θ-регуляризаторов. В то время как Θ-регуляризатор требует, чтобы тематические распределения у близких документов были близкими, Ф-регуляризатор требует, чтобы наличие ссылок на близкие документы несло сходную информацию.

Это означает, что Ф-регуляризатор напрямую не учитывает, что метка $[v]$ и документ v как-то связаны. Возможна странная ситуация, когда

для какой-то темы t справедливо, что $\phi_{[v]t} = 1$ (если автор пишет о теме t, то он обязательно цитирует статью v), но $\theta_{vt} = 0$ (статья v никак не затрагивает тему t).

Существуют два способа принять во внимание эту связь: корректировка $\phi_{[v]t}$ с учётом θ_{vt}, либо корректировка θ_{vt} с учётом $\phi_{[v]t}$. Архитектура пакетного онлайнового EM-алгоритма позволяет эффективно реализовать лишь последний вариант.

Введём "гибридный" регуляризатор, зависящий от Φ и Θ и оценивающий кросс-энтропию разрыва между $p(t \mid v)$ и $p(t \mid [v])$:

$$R(\Theta, \Phi) = \sum_d \sum_t \frac{n_t}{n_{[d]}} \phi_{[d]t} \log \theta_{td}$$

Ему соответствует следующий M-шаг:

$$\theta_{td} = \underset{t}{\text{norm}}(n_{dt} + \sum_t \frac{n_t}{n_{[d]}} \phi_{[d]t})$$

Эксперименты показывают, что гибридный регуляризатор лучше всего работает в сочетании с другими регуляризаторами.

5.5 Обсуждение кросс-энтропийного регуляризатора

В этом разделе будут приведены соображения общего характера, касающиесся регуляризатора "ссылка в теме" и регуляризатора "тема с учётом ссылки".

Пусть даны два распределения вероятностей p, q и необходимо изменить их таким образом, чтобы уменьшить кросс-энтропию $H(p, q) = -\sum_x p(x) \log q(x)$.

Нетрудно показать, что $H(p, q) = H(p) + D_{KL}(p||q)$.

Если зафиксировать p и вести минимизацию по q, то слагаемое $H(p)$ становится константой. Отсюда имеем, что данная задача эквивалентна минимизации дивергенции Кульбака-Лейблера. В силу её свойств получаем, что оптимум достигается при $q = p$. Этого можно достичь, например, "смешивая" распределения: $q^{(t)} = \underset{x}{\text{norm}}\, q^{(t-1)} + p$.

Если зафиксировать q и вести минимизацию по p, то оптимум не обязательно достигается в точке $q = p$. Если для какого-то x_0 вероятность $q(x_0)$ очень мала, то $\log q(x_0)$ есть огромное по модулю число. Поэтому может оказаться выгоднее не устанавливать $p(x_0) = q(x_0)$, а положить $p(x_0) = 0$, полностью обнулив слагаемое $\log q(x)$.

Формулы M-шага регуляризаторов для Φ можно проинтерпретировать с точки зрения этих двух стратегий. Заметим, что все формулы содержат два слагаемых, отвечающих за исходящие и входящие ссылки соответственно.

Первое слагаемое просто приближает распределение в d к усреднённому распределению в соседях d. Это соответствует минизации кросс-энтропии по q.

Второе слагаемое соответствует минизации по p. Рассмотрим его действие подробно.

Пусть некий документ v ссылается на документ d, причём $\phi_{[v]t}$ близко к нулю: тогда данное преобразование будет стремиться обнулить $\phi_{[d]t}$. Более того, это приводит к цепной реакции: если в итоге значение $\phi_{[d]t}$ обнулится (или просто существенно уменьшится), то точно также обнулятся и $\phi_{[u]t}$ для всех u, на которые ссылается d. Таким образом, данное слагаемое приводит к разреживанию ссылочной модальности матрицы Φ.

Эксперименты показали, что при больших τ вклад первого слагаемого адекватен (не приводит к сильному ухудшению качества; в ряде случаев оптимальное значение $\tau \sim 10^7$), но при этом использование большого коэффициента регуляризации для второго слагаемого быстро приводит к полному занулению всех величин $\phi_{[d]t}$ и большой численной неустойчивости алгоритма.

Это означает, что использование одинакового коэффициента регуляризации для обоих слагаемых нецелесообразно. С точки зрения полувероятностного подхода APTM в этом и нет необходимости: каждое слагаемое можно трактовать как отдельную регуляризационную добавку; можно даже совмещать регуляризационные добавки от разных способов вывода (например, сложив градиентную оптимизацию входящих ссылок

с точки зрения $p(t \mid [v])$ и оптимизацию исходящих ссылок по Φ, построенную на применении условий Каруша-Куна-Таккера, или даже результат применения какого-то эвристического подхода).

В работе [13] было принято решение отказаться от разреживающей регуляризационной добавки, но этому не было приведено никаких обоснований. В данной работе мы поступим аналогично, но изучим её действие в рамках дополнительного эксперимента.

6 Эксперименты

6.1 Методология

Я взял два регуляризатора из литературы (netPLSA и iTopic) и реализовал их на базе своей реализации PLSA. Также я реализовал учёт графовой структуры посредством ссылочной модальности, одну из вариаций кросс-энтропийного регуляризатора для ссылочной модальности и их комбинацию (на основании этой реализации я произвёл перекрёстную сверку результатов, чтобы убедиться в отсутствии программных недочётов).

Кроме того, я реализовал все вышеперечисленные регуляризаторы в рамках открытой библиотеки bigARTM.

Далее я сравнивал эти алгоритмы по различным показателям и рассматривал варианты их улучшений.

Проводилось 250 итераций, каждая итерация bigARTM состояла из пяти внутренних. Вес модальности ссылок γ был установлен в 0.5.

6.2 Данные

Из архива DBLP были взяты сведения о всех статьях, опубликованных на четырёх конференциях (WWW, KDD, NIPS, SIGIR) до 2009 года включительно.

Вершина графа - это автор (конкатенация названий всех его статей), между вершинами проводится ребро за каждую статью, написанную в соавторстве. Метка автора - номер конференции, внутри которой у него наибольшее число публикаций.

В корпусе 13770 авторов, 56540 рёбер и 10113 документов.

6.3 Метрики

(здесь и далее, если не указано обратного, то оценивается не сама матрица Θ, а "жесткая" классификация вершин, полученная из неё взятием номера наиболее вероятной темы)

Поскольку ранее каждому автору было сопоставлена одна из четырёх конференций, то имеется образец, в сравнении с которым возможно измерить качество классификации.

- Normalized Mutual Information: показывает количество информации, содержащейся о "правильных" метках конференций. Величина лежит в $[0, 1]$, чем больше, тем лучше..

- Точность классификации: если использовать тематику документа, как признаки для логистической регрессии, решающей задачу угадывания "правильных" меток конференций, то точность классификации также будет метрикой качества тематической модели..

В то же время, алгоритм может находить какие-то другие естественные сообщества, не обязательно отображающиесся на конференции. Это требует метрик, оценивающих кластеризацию "саму по себе".

- Вес разреза: сумма весов межкластерных рёбер. Подразумевается, что между разными кластерами должно быть мало связей.

- Ratio Cut: вес разреза, нормированный на число вершин внутри кластера. Ведёт к более сбалансированным кластерам. Минимизируется.

- Normalized Cut: вес разреза, нормированный на общий вес рёбер, хотя бы один конец которых находится внутри кластера. Эта метрика штрафует за вынесение изолированных вершин в отдельный кластер. Минимизируется.

- Модулярность: мера того, насколько число "внутренних" рёбер отличается от ожидаемого, если рёбра раскидывались бы случайным образом. Максимизируется.

Также можно рассматривать и значение функции, которую оптимизируют регуляризаторы. Эта величина оценивается непосредственно по распределению Θ, а не по его наиболее вероятным элементам.

Кроме того, были предприняты попытки оценки когерентности [9] и PMI, но эти метрики оказались сложными для интерпретации.

6.4 Начальное приближение

Известно, что результат задачи тематического моделирования существенно зависит от начального приближения матриц Φ и Θ. Широко используется метод инициализации, в котором матрица Φ заполняется случайными значениями и нормируется, а матрица Θ инициализируется равномерным распределением.

Если рассматривать постановку мультимодальной задачи, то необходимо подобрать начальное приближение уже для трёх матриц: Θ, Φ и Φ^2.

Я провёл исследование и установил, что начальное приближение Φ^2 является очень важным фактором. Использование случайного начального приближения ведёт к результатам, значительно худшим, чем использование равномерного начального приближения.

У происходящего есть разумная интерпретация. Случайная инициализация Φ полезна для того, чтобы разорвать начальную симметрию и первоначально выделить какие-то слова в какие-то темы. В то же время кажется, что нет причин заранее выделять ссылки в какие-то темы: разумнее определять значение ссылки за счёт тематических распределений в документах, с которыми она связана.

На этом графике изображено следующее:

- Разноцветные пунктирные линии — это комбинированный регуляризатор с различными γ, запущенный на равномерном начальном приближении Φ^2.
- Разноцветные пунктирные линии с точками — аналогично, но на случайном Φ^2.
- Регуляризатор лог-правдоподобия ссылочной модальности — сплошная чёрная линия для равномерного Φ^2 и чёрная линия с точками для случайного Φ^2
- BigARTM со ссылочной модальностью — сплошная жёлтая линия с точками для случайного Φ^2 и без точек для равномерного

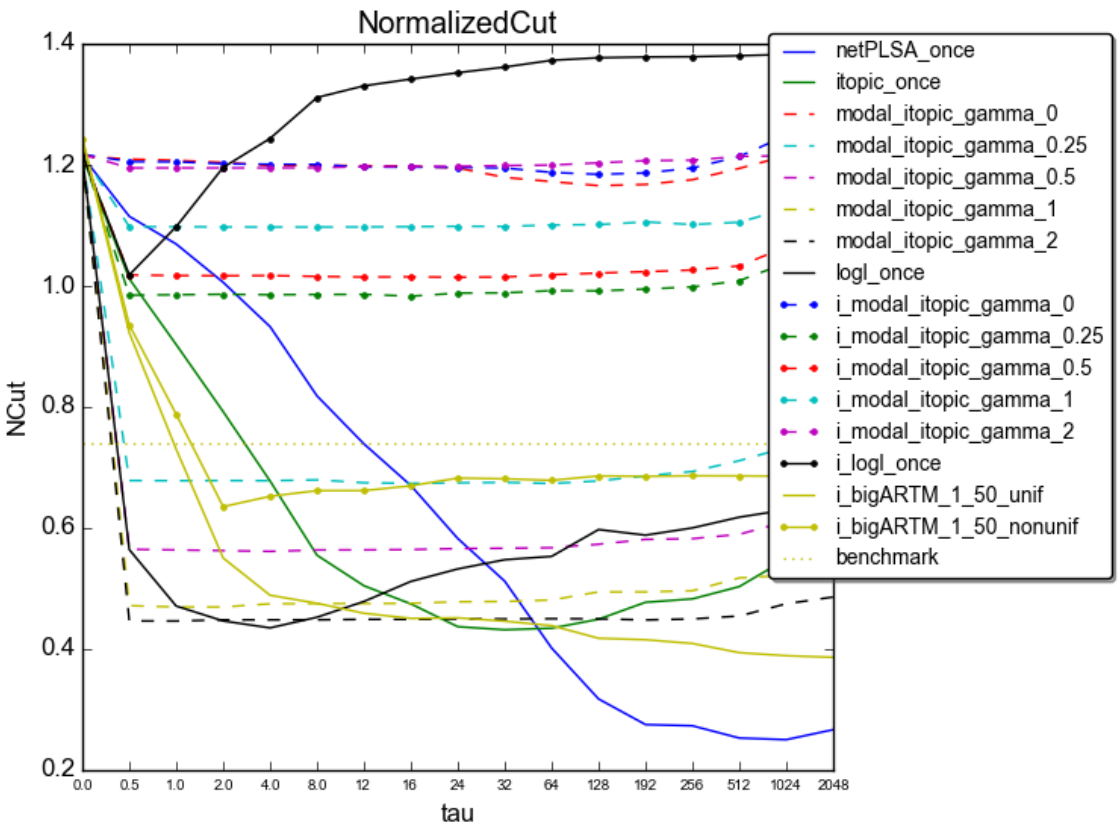


Рис. 3: Метрика NCut в зависимости от значения τ для различных регуляризаторов и начальных приближений

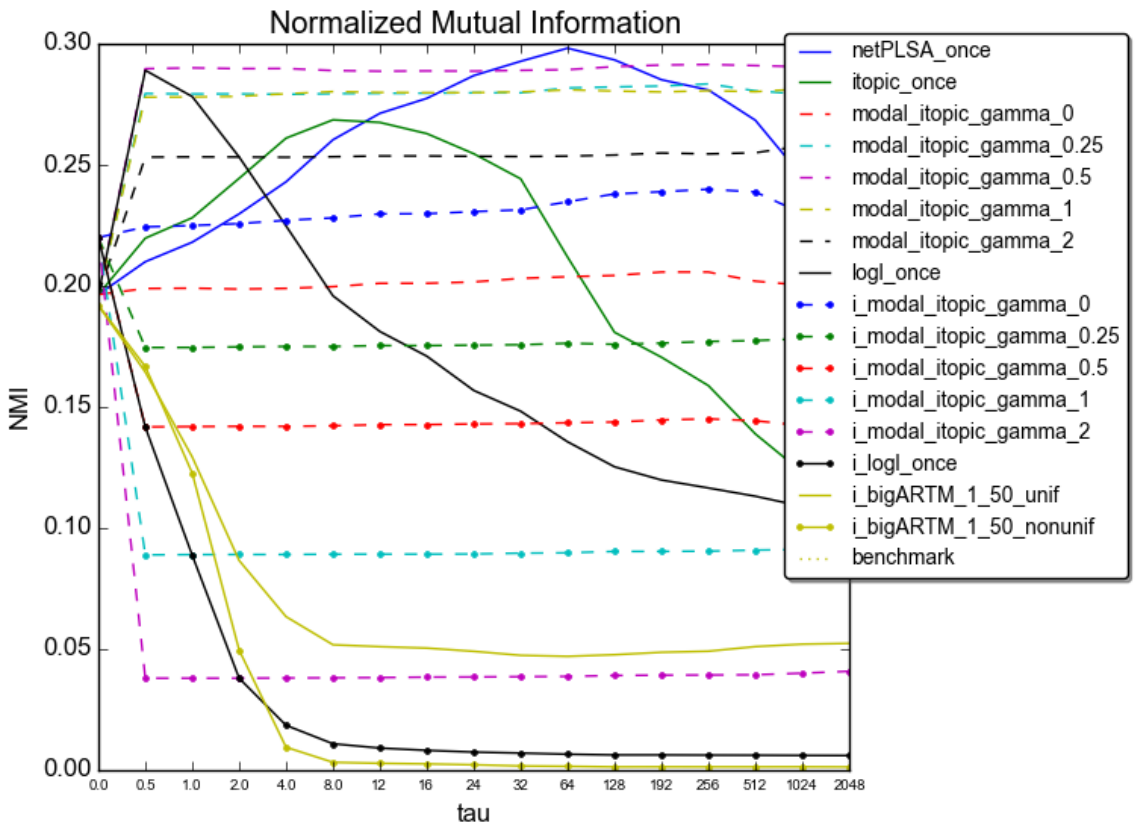


Рис. 4: Метрика NMI в зависимости от значения τ для различных регуляризаторов и начальных приближений

Виден огромный разрыв между двумя инициализациями для ссылочной модальности.

Виден значительный разрыв между комбинированными регуляризаторами (исключая линию с $\gamma = 0$, которая находится вместе с "точечными" линиями).

Виден менее очевидный разрыв между двумя инициализациями bigARTM, растущий с ростом τ.

Все кластерные метрики выглядят аналогично и также наглядно показывают, что инициализация равномерным Φ^2 значительно лучше.

График NMI не настолько прозрачен (разрыв между линиями меньше и они частично накладываются), но общая ситуация совершенно аналогична.

6.5 Эффект от внутренних итераций M-шага

Можно рассмотреть две модификации netPLSA (см. листинг 1): изначальная и та, в которой происходит лишь одна итерация M-шага. Назовём их `netPLSA until` и `netPLSA once` соответственно.

По аналогии с netPLSA можно построить модификацию iTopic (`iTopic until`), где регуляризаторная добавка прибавляется несколько раз в цикле.

Эксперимент показывает, что until-модификации показывают лучшее качество (но не драматически), но при этом вычисляются в несколько раз дольше.

С другой стороны, until-модификации (особенно netPLSA) не показывают чёткого оптимума по τ. Можно сделать предположение о том, что он может достигаться при $\tau > 2048$. В следующем разделе будет показано, что это скорее всего не так.

6.6 Анализ внутренних итераций M-шага

Назовём *обновлением* один пересчёт матрицы Θ (иными словами, пересчёт Θ по формулам для PLSA либо изменение под действием регуляризатора).

Вспомним, что в EM алгоритме есть целевой функционал $Q = Q_{\text{PLSA}} + \tau Q_R$, который оптимизируется на каждой итерации. Первое слагаемое связано с правдоподобием: $Q_{\text{PLSA}} = \sum_d \sum_w n_{dw} \sum_t p_{tdw} \log[\phi_{wt}\theta_{td}$. Второе слагаемое связано с регуляризатором.

Для until-модфикаций цепочка обновлений выглядит так:

1. Обновляются θ_{td}. Это повышает Q_{PLSA}, но может понизить Q_R. Суммарно это может привести к понижению Q. Это может зависеть от

2. Происходит некоторое количество итераций регуляризатора, который повышает Q_R и, возможно, понижает Q_{PLSA}. Итерации прекращаются в момент, когда:

   - превышено их максимальное число
   - от них перестал наблюдаться суммарный положительный эффект
   - падение Q в первом пункте полностью компенсировано

3. Увеличивается счётчик итераций, возвращение к пункту 1

Можно нарисовать графики Q_{PLSA} и Q_R в зависимости от номера апдейта и отметить на них жирными точками моменты, когда цикл внутренних итераций заканчивался.

Быстро сходящемусся итерационному процессу на таком графике будет соответствовать вытянутая вверх "кобра". Итерационный процесс, в котором происходит сильная конкуренция шагов по оптимизации Q_{PLSA} и Q_R будет изображён в виде длинной вытянутой "пилы".

Это даёт понимание внутренней работы алгоритма. Например, это позволяет подобрать максимальное число итераций (увидеть порог, после которого увеличение Q резко замедляется).

Далее приведены графики обновлений для netPLSA until и iTopic until.

Видно, что при небольших τ внутренние итерации iTopic until заканчиваются "досрочно", а при больших — всегда заканчиваются по усло-

вию максимума итераций. Также заметно, что разные значения τ ведут к ощутимо различным траекториям.

Видно, что амплитуда колебаний netPLSA until велика по сравнению с величиной, на которую выросло Q.

Вероятно, это обьясняет выход на плато и показывает, что из плато нельзя выбраться "простыми" мерами наподобие расширения диапазона τ или увеличением числа итераций. Скорее всего, поможет изменение τ в процессе работы, закрепление отобранных тем и какие-то похожие (более тонкие) меры.

Также это показывает, что until-модификации расходуют машинное время очень неэкономно. Единицу вычислительного времени намного выгоднее "вложить" в увеличение числа итераций once-модификации, чем в увеличение числа внутренних итераций M-шага.

В связи с этим в текущей работе в качестве образцов рассматривались только once-модификации.

6.7 Разреживающий кросс-энтропийный регуляризатор

Пусть γ — вес модальности ссылок, также зафиксируем какое-то натуральное число P. Рассмотрим следующую формулу для M-шага:

$$\phi_{[d]t} = \underset{[d]}{\text{norm}} \left(n_{[d]t} + \tau r_{[d]t} + \beta r'_{[d]t} \right)$$

Слагаемое $r_{[d]t}$ соответствует сглаживающей части кросс-энтропийного регуляризатора. В случае регуляризации распределения $p(t \mid [v])$ оно равняется $\sum_{d,v} w(d,v) \frac{p(t|[v])}{n_{[d]t}}$. Коэффицент $\tau = 1$.

Слагаемое $r_{[d]t}$ отвечает за разреживание ссылочной модальности матрицы Ф:

$$r'_{[d]t} = \frac{1}{n_{[d]}} \sum_{u,d} w(u,d) \log(p(t \mid [u])) = \frac{1}{n_{[d]}} \sum_{u,d} w(u,d) \log \frac{n_{[u]t}}{n_{[u]}}$$

Коэффициент β равен 0.2 в течении первых 5 итераций, затем обнуляется в течении следующих P итераций, затем снова устанавливается

в 0.2 на ближайшие 5 итераций и так далее (иными словами, $\beta(i) = 0.2 \cdot [(i\%P) < 5]$). Решение о периодическом включении и выключении разреживающего регуляризатора связано с тем, что иначе он действует слишком агрессивно.

Таблица 1: Результаты эксперимента с разреживающим регуляризатором (250 итераций, 5 внутренний итераций, среднее по 4 перезапускам)

NCut		–	P = 10	P = 25	P = 50	P = 100
`p(t|[v])`	`(gamma = 0.5)`	0.6641	0.7698	0.6689	0.6465	**0.6421**
`p(t|[v])`	`(gamma = 1)`	0.5322	0.6883	0.5509	0.5378	**0.5293**
NMI		–	P = 10	P = 25	P = 50	P = 100
`p(t|[v])`	`(gamma = 0.5)`	**0.4222**	0.3915	0.4108	0.4127	0.4099
`p(t|[v])`	`(gamma = 1)`	0.3781	0.3670	0.3840	0.3860	**0.3866**

Эксперимент показал, что разреживание улучшает качество выделения кластеров, но неоднозначно взаимодействует с NMI: улучшает его при $\gamma = 1$, но ухудшает при $\gamma = 0.5$, причём эффект изменения γ оказывается сильнее, чем эффект введения разреживания.

Эти эксперименты демонстрируют относительную пользу разреживающего кросс-энтропийного регуляризатора, одновременно говоря о необходимости более тонкой настройки параметров.

6.8 Сравнение с образцами

На графиках изображены:

- Чистый PLSA без учёта графа (в реализации на основе bigARTM)
- Регуляризатор iTopic для двух значений τ
- Регуляризатор лог-правдоподобия ссылочной модальности
- Комбинация гибридного регуляризатора и кросс-энтропийного регуляризатора на основе $p(t \mid [v])$
- Кросс-энтропийный регуляризатор на основе ϕ

Метрика	NCut (лучше меньше)	NMI (лучше больше)
`iTopic (tau=4)`	**0.519303826**	0.361636097
`netPLSA (tau=32)`	0.421937332909	0.365562190318
`PLSA`	0.836377842	**0.393541527**
`modal likelihood`	0.619220682	0.350554135
`phi (tau = 1024)`	0.526113909	0.378951116
`phi + hybrid`	**0.309823317**	0.255207926
`p(t|[v]) (tau=1)`	0.661924108	**0.414944947**
`p(t|[v]) + hybrid`	0.6377828	**0.39995478**

Коэффициент перед регуляризатором лог-правдоподобия ссылочной модальности $\gamma = 0.5$, коэффициент перед гибридным регуляризатором $\beta = 1$. В bigARTM использовалось 5 внутренних итераций.

Эксперименты показывают, что введение регуляризаторов позволяет существенно улучшить как результат `PLSA` (который не учитывал никакую информацию о графе), так и "базового" варианта `modal likelihood` (где была только модальность ссылок, но не было никакой дополнительной регуляризации).

Модель `phi (tau=1024)` показывает качество, аналогичное образцу `iTopic (tau=4)`. Ряд моделей улучшает каждую из метрик по отдельности, но пока что не получилось построить модель, показывающую существенно лучшее качество по всем метрикам вместе, что является естественным направлением для дальнейшей работы.

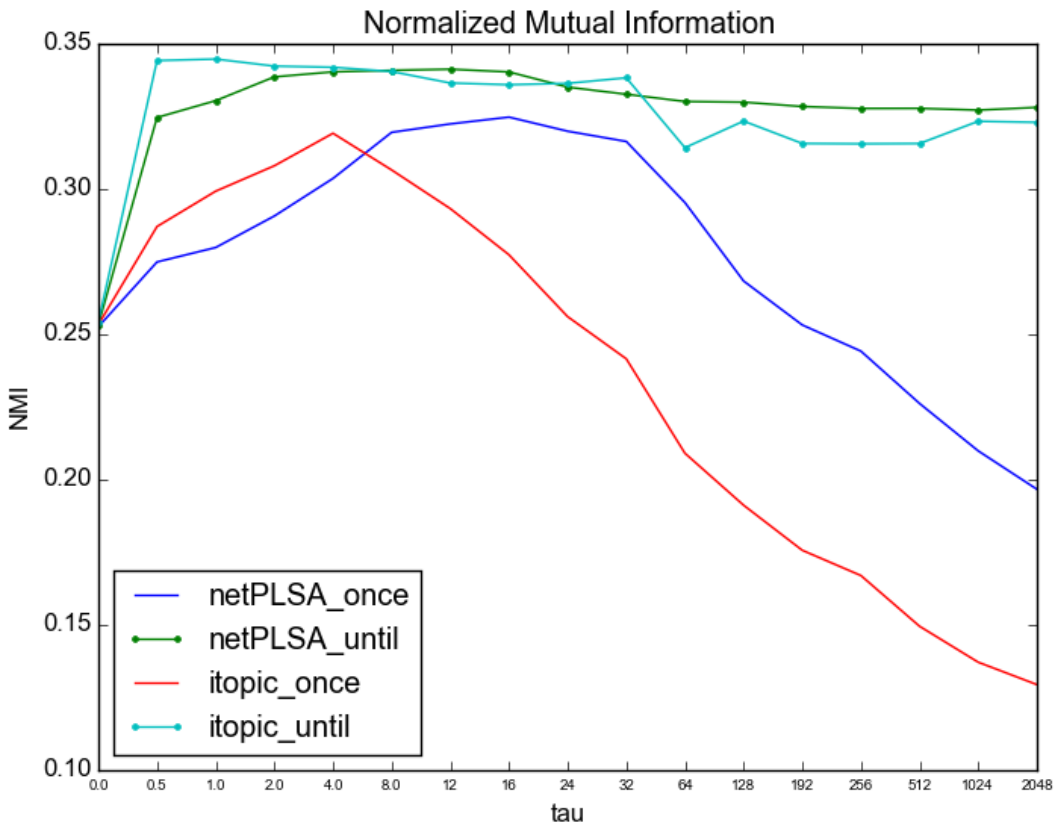
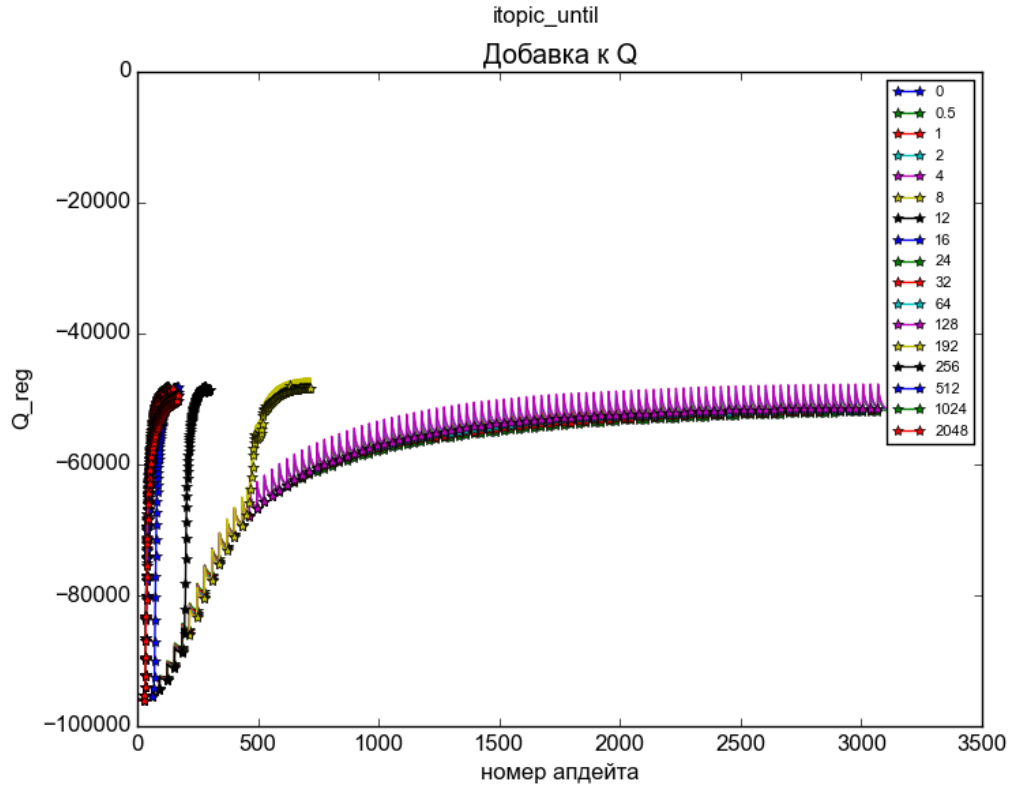
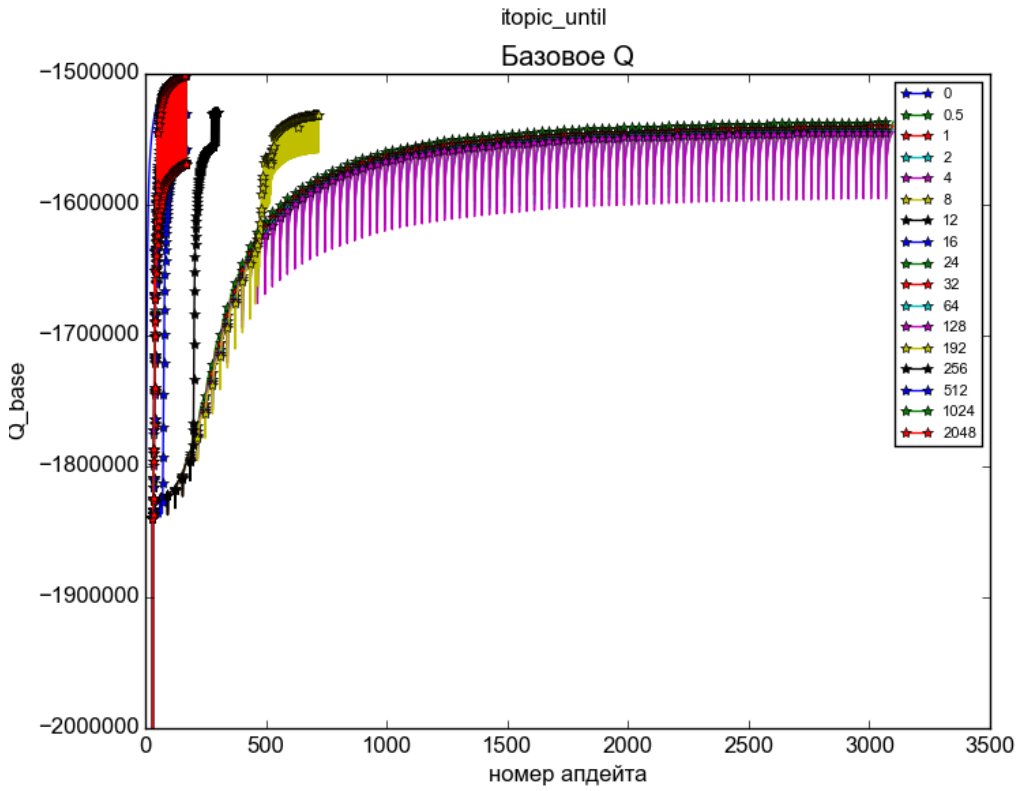


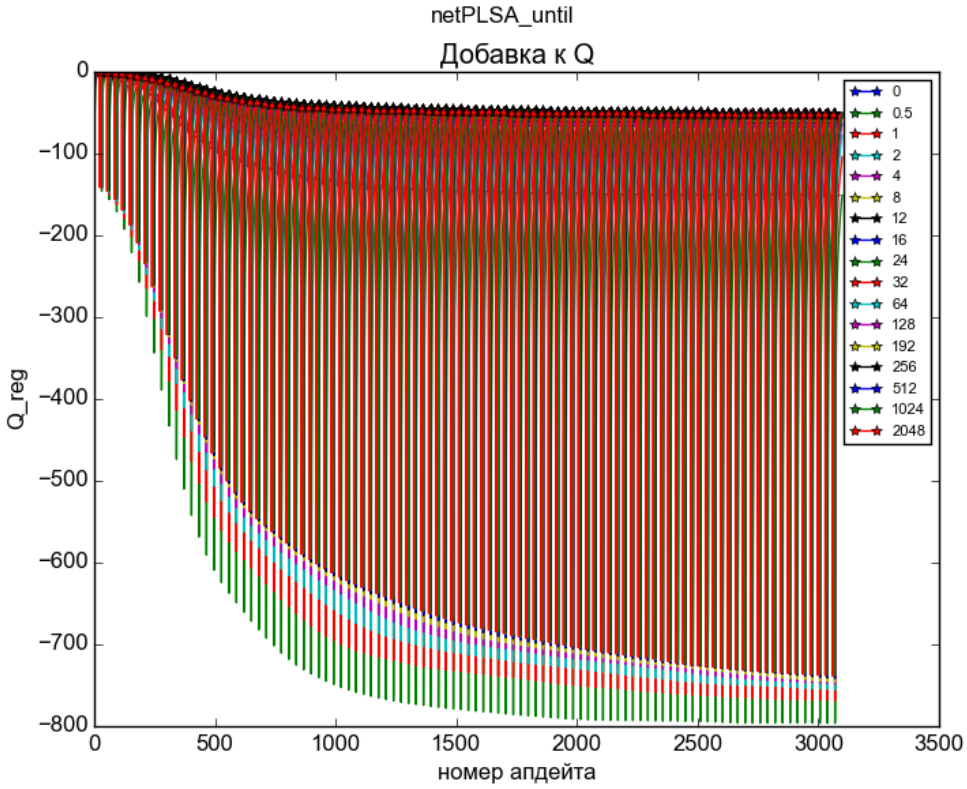
Рис. 5: Сравнение until- и once-модификаций для разных τ, проведено 50 итераций EM-алгоритма



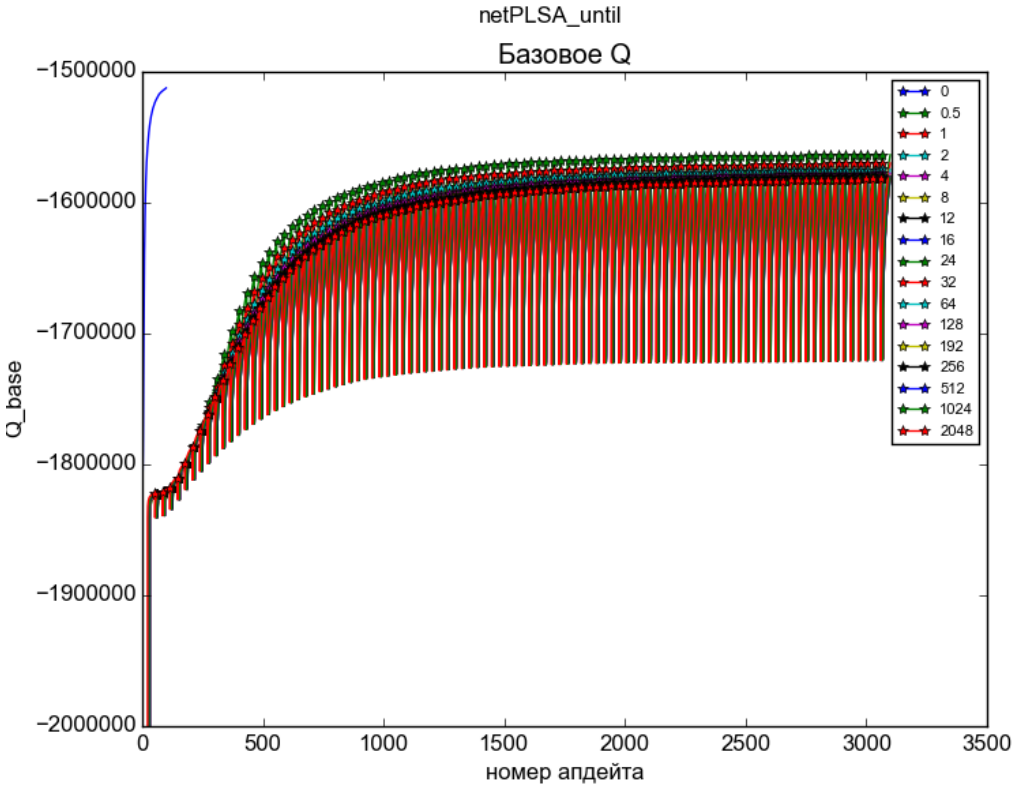
(a) Значение базового Q для iTopic until



(b) Значение дополнительного Q для iTopic until



(a) Значение базового Q для netPLSA until



(b) Значение дополнительного Q для netPLSA until

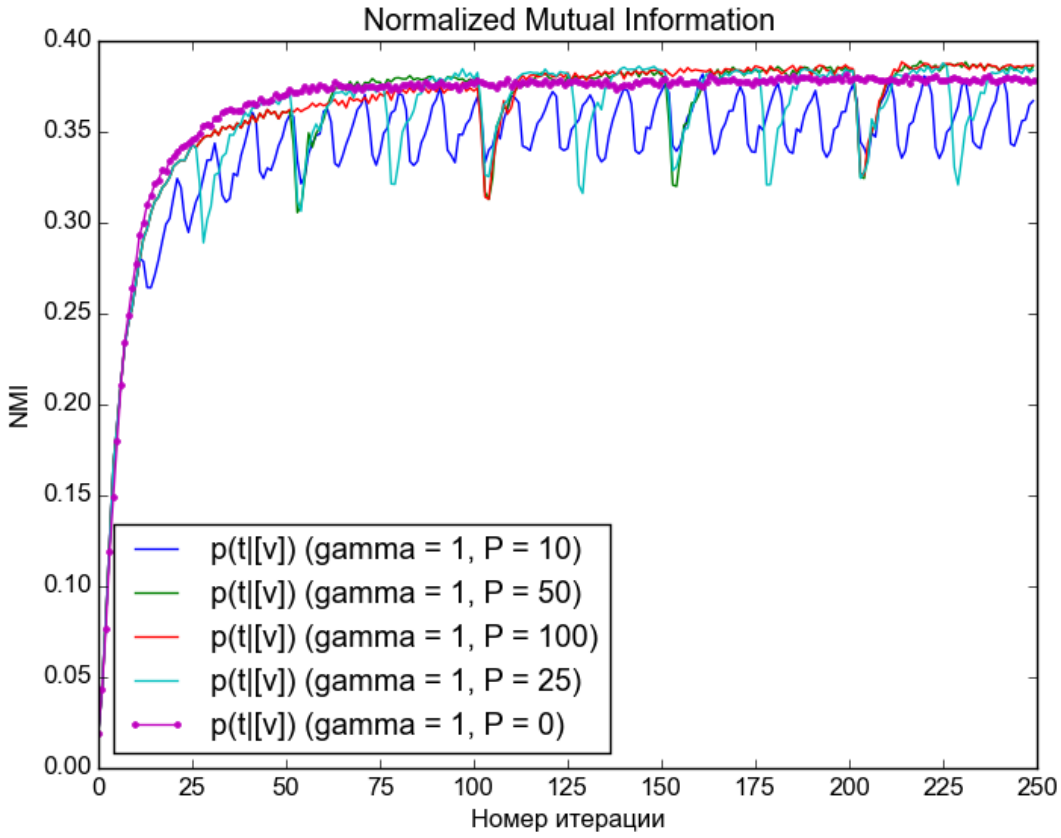


Рис. 6: Эффект разреживающего регуляризатора на NMI

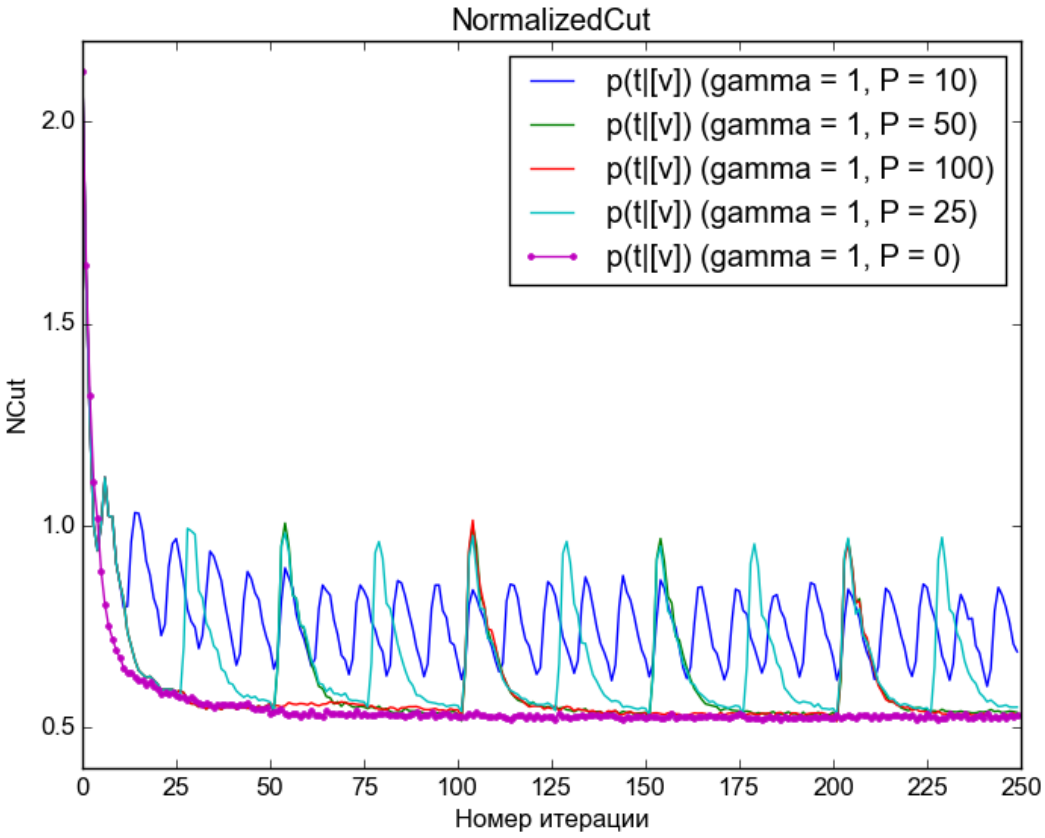


Рис. 7: Эффект разреживающего регуляризатора на NCut

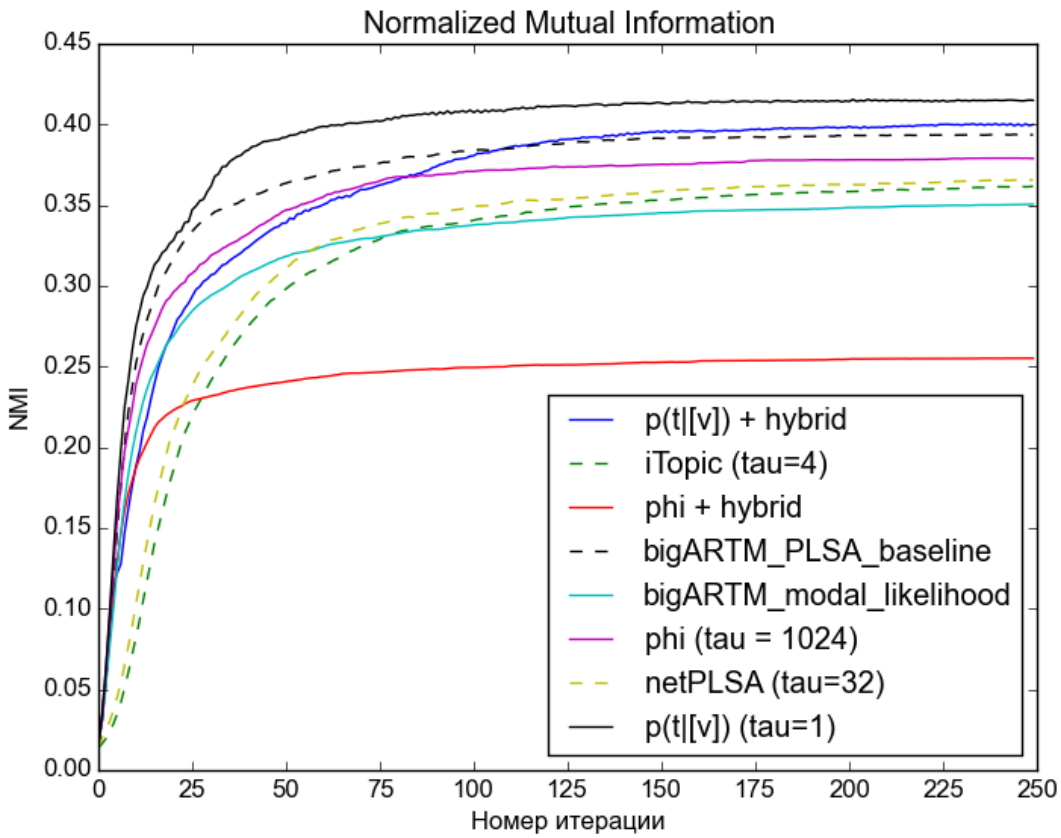


Рис. 8: Сравнение нескольких моделей с образцами для метрики NMI. 250 итераций, среднее по 4м перезапускам

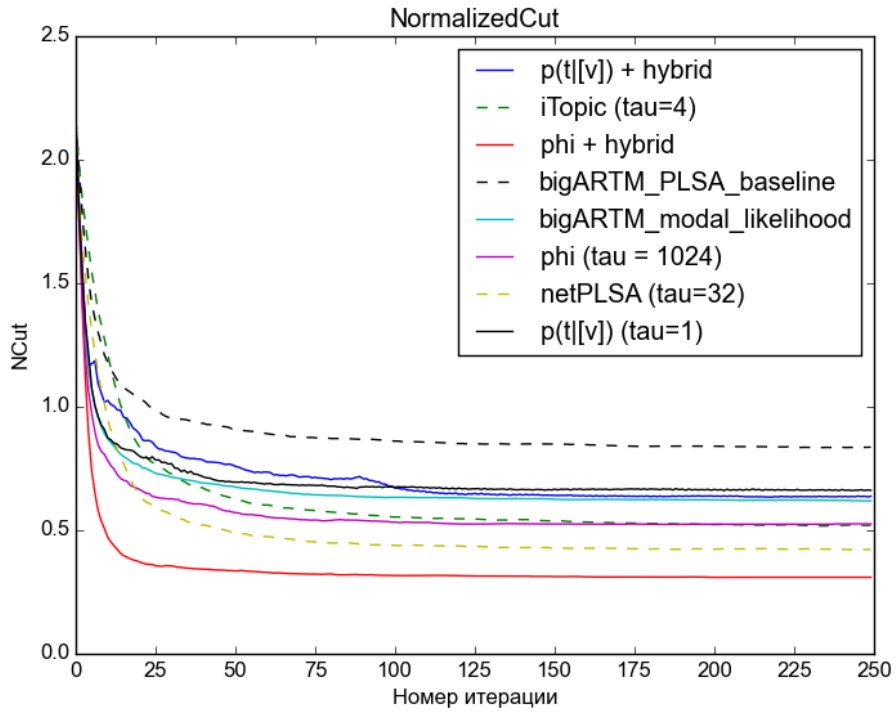


Рис. 9: Сравнение нескольких моделей с образцами для метрики NCut. 250 итераций, среднее по 4м перезапускам

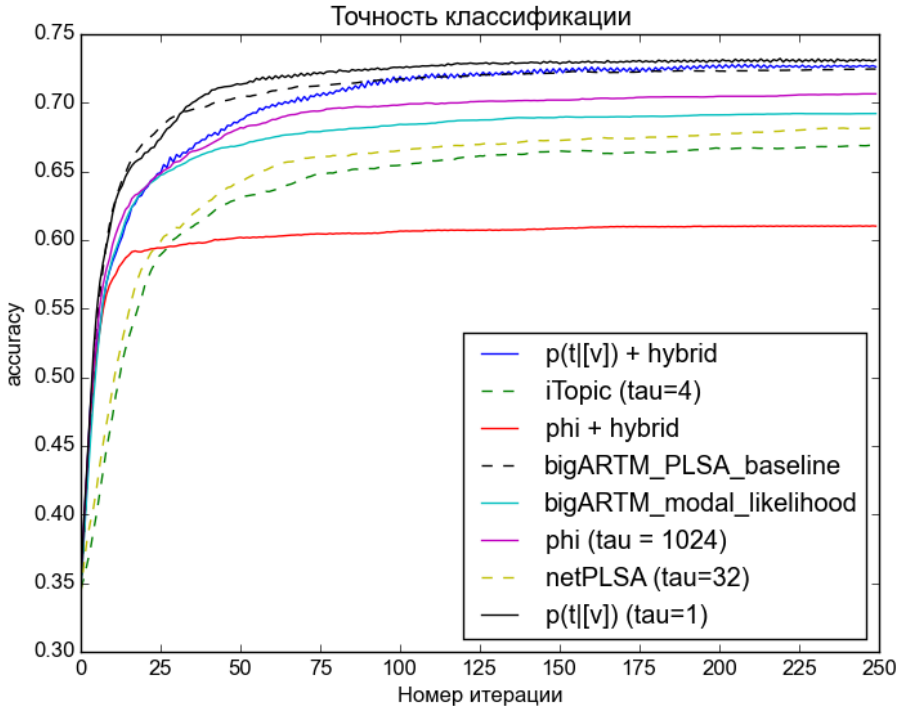


Рис. 10: Сравнение нескольких моделей с образцами для метрики точности классификации. 250 итераций, среднее по 4м перезапускам

7 Приложение

7.1 Отношения правдоподобия

Величина $p(t \mid [v])$ — апостериорная вероятность темы t с учётом наличия ссылки на v внутри документа. Эту вероятность можно вычислить по формуле Байеса, зная априорную вероятность $p(t)$, правдоподобие $p([v] \mid t)$, а также нормировочную константу в знаменателе.

С другой стороны, искомую вероятность можно найти, зная лишь два числа: априорную вероятность $p(t)$ и величину $\lambda = \frac{p([v]|t)}{p([v]|\neg t)}$, называемую *отношением правдоподобия*:

$$p(t \mid [v]) = \frac{p(t)p([v] \mid t)}{p(t)p([v] \mid t) + (1 - p(t))p([v] \mid \neg t)} = \frac{p(t)}{p(t) + (1 - p(t))\lambda}$$

Такая форма записи подчёркивает, что результат пересчёта вероятностей по формуле Байеса определяется в первую очередь отношением правдоподобия.

Ещё более наглядно это иллюстрирует "шансовая форма" формулы Байеса. Сопоставим событию X не его вероятность $P(X)$, а его шансы $O(X) = \frac{P(X)}{P(\neg X)}$. Между вероятностями из интервала $(0, 1)$ и шансами из $(-\infty, +\infty)$ существует биекция, определяемая соотношением $o(x) = \frac{p(x)}{1-p(x)}$.

Аналогично вводятся апостериорные шансы X с учётом события E:

$$O(X \mid E) = \frac{P(X \mid E)}{P(\neg X \mid E)}$$

Записав формулу Байеса для $P(X \mid E)$, для $P(\neg X \mid E)$, а затем разделив одно на другое, получим:

$$O(X \mid E) = O(X) \cdot \frac{P(E \mid X)}{P(E \mid \neg X)} = O(X) \cdot \lambda(E \mid X)$$

Видно, что апостериорные шансы представляют из собя просто априорные шансы, умноженные на отношение правдоподобия.

Более того, для независимых в совокупности событий E_i (а в рамках вероятностной модели PLSA метки всех слов и объектов других модальностей являются условно независимыми с учётом известной темы) справедливо очень удобное соотношение:

$$O(X \mid E_1, E_2, \dots E_k) = O(X) \cdot \prod_i \lambda(E_i \mid X)$$

Таким образом, отношение правдоподобия можно проинтерпретировать, как величину, показывающую, насколько нужно сместить свою оценку в свете новой информации.

Поэтому рассуждения вида «*наличие ссылки на документ A и наличие ссылки на документ B должны нести схожую информацию*» очень естественно переносятся на отношения правдоподобия.

Теперь распишем отношение правдоподобия для произвольного токена w:

$$\lambda(w \mid t) = \frac{p(w \mid t)}{p(w \mid \neg t)} = \frac{\phi_{wt}}{\sum_{k \in T, k \neq t} p(k \mid \neg t) p(w \mid k, \neg t)} =$$

$$\frac{\phi_{wt}}{\sum_{k \in T, k \neq t} \frac{p(k)}{p(\neg t)} \phi_{wk}} = \frac{\phi_{wt} \cdot \sum_{k \in T, k \neq t} p(k)}{\sum_{k \in T, k \neq t} p(k) \phi_{wk}}$$

Если сделать допущение о том, что все пропорции тем одинаковы, $p(k) = \rho = \text{const}$, то можно получить следующую приближенную формулу:

$$\lambda(w \mid t) \approx \frac{\phi_{wt} \cdot (T-1)\rho}{\rho \sum_{k \in T, k \neq t} \phi_{wk}} \propto \frac{\phi_{wt}}{\sum_{k \in T, k \neq t} \phi_{wk}}$$

Можно показать, что с учётом ограничения нормировки набор $\{\lambda(w \mid t) \mid t \in T\}$ однозначно определяет набор $\{\phi_{wt} \mid t \in T\}$. Следовательно, регуляризуя Φ мы неявно регуляризуем λ и наоборот.

Это означает, что регуляризацию матрицы Φ напрямую можно проинтерпретировать в терминах отношения правдоподобия, и такая регуляризация представляется разумной.

Заметим, что величины $p(k)$ можно выразить через частотные оценки $\frac{n_k}{n}$. Это говорит о возможности регуляризации отношений правдопо-

добия напрямую при помощи некой функции $R(\Phi, n_t)$. Это идея представляет собой интересное направление для дальнейшей работы, причём применимость этой идеи не ограничивается рамками учёта графовой структуры.

8 Заключение

В данной работе было предложено несколько аддитивно регуляризованных моделей, учитывающих графовую структуру связей между документами. Модели основаны на регуляризации распределения $p(t \mid [v])$ и/или отношений правдоподобия — понятиях, которые до текущего момента широко не использовались в рамках тематического моделирования.

Предложенные модели удобно комбинировать друг с другом или с другими моделями из обширного списка регуляризаторов ARTM; кроме того, они допускают эффективное распараллеливание.

По ряду метрик эти модели показывают лучшее качество, чем образцы.

Кроме того, данная работа расширяет границы применимости подхода ARTM на случай регуляризаторов, зависящих от ненормированной матрицы Ф.

Список литературы

[1] Jeff A Bilmes et al. A gentle tutorial of the em algorithm and its application to parameter estimation for gaussian mixture and hidden markov models. *International Computer Science Institute*, 4(510):126, 1998.

[2] Jonathan Chang and David M Blei. Relational topic models for document networks. In *AIStats*, volume 9, pages 81–88, 2009.

[3] David Cohn and Thomas Hofmann. The missing link-a probabilistic model of document content and hypertext connectivity. *Advances in neural information processing systems*, pages 430–436, 2001.

[4] Laura Dietz, Steffen Bickel, and Tobias Scheffer. Unsupervised prediction of citation influences. In *Proceedings of the 24th international conference on Machine learning*, pages 233–240. ACM, 2007.

[5] Dongsheng Duan, Yuhua Li, Ruixuan Li, Rui Zhang, Xiwu Gu, and Kunmei Wen. Limtopic: A framework of incorporating link based importance into topic modeling. *IEEE Transactions on Knowledge and Data Engineering*, 26(10):2493–2506, 2014.

[6] Yue Lu, Qiaozhu Mei, and ChengXiang Zhai. Investigating task performance of probabilistic topic models: an empirical study of plsa and lda. *Information Retrieval*, 14(2):178–203, 2011.

[7] Andrew Mccallum, David M Mimno, and Hanna M Wallach. Rethinking lda: why priors matter. In *Advances in neural information processing systems*, pages 1973–1981, 2009.

[8] Qiaozhu Mei, Deng Cai, Duo Zhang, and ChengXiang Zhai. Topic modeling with network regularization. In *Proceedings of the 17th International Conference on World Wide Web*, WWW '08, pages 101–110, New York, NY, USA, 2008. ACM.

[9] David Mimno, Hanna M Wallach, Edmund Talley, Miriam Leenders, and Andrew McCallum. Optimizing semantic coherence in topic models.

In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 262–272. Association for Computational Linguistics, 2011.

[10] Tom M Mitchell et al. Machine learning. wcb, 1997.

[11] David Newman, Sarvnaz Karimi, and Lawrence Cavedon. External evaluation of topic models. In *in Australasian Doc. Comp. Symp., 2009*. Citeseer, 2009.

[12] Mark EJ Newman and Michelle Girvan. Finding and evaluating community structure in networks. *Physical review E*, 69(2):026113, 2004.

[13] Y. Sun, J. Han, J. Gao, and Y. Yu. itopicmodel: Information network-integrated topic modeling. In *2009 Ninth IEEE International Conference on Data Mining*, pages 493–502, Dec 2009.

[14] Konstantin Vorontsov, Oleksandr Frei, Murat Apishev, Peter Romov, and Marina Dudarenko. Bigartm: Open source library for regularized multimodal topic modeling of large collections. In *Analysis of Images, Social Networks and Texts*, pages 370–381. Springer, 2015.

[15] Konstantin Vorontsov and Anna Potapenko. Tutorial on probabilistic topic modeling: Additive regularization for stochastic matrix factorization. In *International Conference on Analysis of Images, Social Networks and Texts_ x000D_*, pages 29–46. Springer, 2014.

[16] Konstantin Vorontsov and Anna Potapenko. Additive regularization of topic models. *Machine Learning*, 101(1-3):303–323, 2015.

[17] Xin Wayne Zhao, Jinpeng Wang, Yulan He, Jian-Yun Nie, and Xiaoming Li. Originator or propagator?: incorporating social role theory into topic models for twitter content analysis. In *Proceedings of the 22nd ACM international conference on Conference on information & knowledge management*, pages 1649–1654. ACM, 2013.

[18] Guoqing Zheng, Jinwen Guo, Lichun Yang, Shengliang Xu, Shenghua Bao, Zhong Su, Dingyi Han, and Yong Yu. Mining topics on

participations for community discovery. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*, pages 445–454. ACM, 2011.

[19] КВ Воронцов. Вероятностное тематическое моделирование. *Курс лекций*, 2013.

[20] Никита Владимирович Дойков. Адаптивная регуляризация вероятностных тематических моделей. *ВКР бакалавра*, 2015.