Полувероятностная тематическая модель для задачи классификации

Смирнов Е. А., Воронцов К. В.
evgenii.smirnov@phystech.edu
Московский физико-технический институт

В работе строится полувероятностная тематическая модель для задачи классификации слов в текстовых документах на основе метода аддитивной регуляризации тематических моделей ARTM с подбором проблемно-ориентированных регуляризаторов. Предложен EM-алгоритм для решения задачи. На основе полученной модели решается задача построения рекомендательной системы для мобильного приложения. Проведён эксперимент на реальных данных мобильного сервиса этого приложения. Сделан вывод о повышении качества рекомендаций сервиса, использующего построенную модель. Новизна заключается в том, что задача классификации слов в текстовых документах сводится к двухматричному разложению вместо трёхматричного.

Ключевые слова: *вероятностное тематическое моделирование, классификация, аддитивная регуляризация тематических моделей ARTM, латентное размещение Дирихле LDA, EM-алгоритм, стохастическое матричное разложение.*

Topic model for classification problem

Smirnov E. A., Vorontsov K. V.

This paper is devoted to topic modeling for classification problem. Model is based on additive regularization of topic models. Offered algorithm for content generation based on this model. It's compared with Gibbs LDA algorithm.Originality of this paper is that model consider data classification.

Keywords: *probabilistic topic modeling, classification,EM-algorithm, additive regularization of topic models, LDA.*

Введение

В работе решается задача построения полувероятностной тематической модели для задачи классификации слов в текстовых документах. Тематическая модель определяет к каким темам и классам относится каждое слово из коллекции документов и какие слова образуют каждую из тем. Вероятностная тематическая модель описывает каждую тему дискретным распределением на множестве тем, каждый документ дискретным распределением на множестве тем и классов. Предполагается, что коллекция документ является последовательностью пар терминов и меток классов, выбранных случайно из смеси этих распределений. Требуется восстановить компоненты смеси по выборке из коллекции документов. Тематические модели для задачи классификации могут быть применены для создания рекомендательных систем.

Доминирующим подходом для построения вероятностных тематических моделей является Латентное размещение Дирихле LDA [1]. Весомым преимуществом LDA является упрощение построения байесовского вывода, что возможно благодаря сопряжённости распределения Дирихле с дискретным распределением. Недостатками этого подхода является отсутствие убедительных лингвистических обоснований, формальная неприемлемость

разреженных решений и трудность построения байесовского вывода, удовлетворяющего большому числу требований к модели.

В данной работе используется альтернативный подход к байесовскому – аддитивная регуляризация тематических моделей ARTM [2], [3]. Построение тематической модели сводится к задаче стохастического матричного разложения. Эта задача в общем случае имеет бесконечное число решений, то есть является некорректно поставленной. ARTM является приложением классической теории регуляризации некорректно поставленных задач к тематическому моделированию. Оно строится на основе модели PLSA[4] с добавлением к логарифму правдоподобия штрафных слагаемых, называемых регуляризаторами. ARTM имеет ряд серьёзных отличий от байесовского подхода. Во-первых, не ставится задача построения чисто вероятностной модели. Что даёт возможность использования оптимизационных критериев вместо априорных распределений для учёта лингвистических ограничений. Очевидность структуры модели отбрасывает надобность в её графическом пояснении. Во-вторых, вместо байевского вывода используется EM-алгоритм. ARTM не требует интегрирования по пространству параметров модели, а сводится к дифференцированию по параметрам. Построение многоцелевых тематических моделей заметно упрощается благодаря аддитивности регуляризаторов. Добавление регуляризаторов приводит к небольшой модификации M-шага. На основе ARTM могут быть построены разные тематические модели в зависимости от выбора вероятностного пространства и вероятностных предположений. Рассматриваемая задача классификации может быть сведена к трёхматричному разложению [2]. В данной работе предлагается двухматричный подход. Его преимуществом является высокая скорость работы, благодаря меньшему числу операции и возможность использования известных регуляризаторов.

Цель работы – построить тематическую модель для задачи классификации слов в текстовых домументах. На её основе построить алгоритм EM-алгоритм. Построить алгоритм рекомендательной системы на основе построенной модели, провести вычислительных эксперимент на реальных данных и оценить его качество.

Постановка задачи тематического моделирования

Пусть D - коллекция текстовых документов, W множество всех употребляемых терминов, содержащихся в документах, C - множество классов заданных на словах из коллекции документов. Будем рассматривать документ $d \in D$ как последовательность пар (w, c), где $w \in W$ – термин, а $c \in C$ – его класс в документе d. Последнее предположение эквивалентно тому, что релевантность термина в разных документах, в общем случае, различна. Появление термина w в документе d будем объяснять некоторой темой $t \in T$. Документы $d \in D$ и термины $w \in W$ являются наблюдаемыми переменными, а темы $t \in T$ скрытыми переменными.

Гипотезы независимости. Будем считать, что не важен порядок документов и порядок терминов в документе. Введем предположение о том, что распределение терминов по темам является общим для всей выборки и не зависит от документа. Естественно предполагать, что каждый документ содержит небольшое число слов из коллекции и к каждому классу относится небольшое число слов.

Описываемый процесс получения пары (w, c) в документе выглядит так

1. Выбирается тема t из распределения $p(t|d)$;
2. Из темы t выбирается слово на основании $p(w|t)$;
3. Класс релевантности c выбирается на основе $p(c|t, d)$, т.е. руководствуясь исключительно темой t.

Эти предположения эквивалентны следующим гипотезам условной независимости:

$$p(w, c|t, d) = p(w|t, d)p(c|t, d)$$

$$p(w|t, d) = p(w|t)$$

Запишем тематическую модель нашей коллекции

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

$$= \sum_{t\in T} p(w|t)p(c, t|d) = \sum_{t\in T} \varphi_{wt}\theta_{ctd}$$

Равенство можно рассматривать, как задачу стохастического разложения матрицы

$$F = (p_{wcd})_{W\times C\times D}$$

в виде произведения $F_c \approx \Phi\Theta_c$ двух неизвестных матриц меньшего размера – матрицы терминов Φ и матрицы тем и классов документов Θ

$$\Phi = (\varphi_{wt})_{W\times T} \qquad\qquad \Theta = (\theta_{ctd})_{C\times T\times D} \tag{1}$$

Параметрами нашей модели соответственно являются φ_{wt} и θ_{ctd}. Восстановив их по выборке, мы можем выбирать термины из определённого класса c^* на основании распределения

$$p(w|c^*, d) = \frac{p(w, c^*|d)}{p(c^*|d)} = \frac{p(w, c^*|d)}{\sum_{t\in T} p(c^*, t|d)} = \frac{\sum_{t\in T} \varphi_{wt}\theta_{c^*td}}{\sum_{t\in T} \theta_{c^*td}}$$

63 Нахождения параметров Φ, Θ эквивалентно максимизации по ним функционала

$$Q(\Phi, \Theta) = \sum_{d\in D}\sum_{w\in W}\sum_{c\in C} n_{dwc} \log(\sum_{t\in T} \varphi_{wt}\theta_{ctd}) + R(\Phi, \Theta),$$

где $R(\Phi, \Theta)$ сумма регуляризаторов. При вероятностных предположениях

$$\sum_{w\in W} \varphi_{wt} = 1,\ \varphi_{wt} \geqslant 0,\ \sum_{c\in C}\sum_{t\in T} \theta_{ctd} = 1,\ \theta_{ctd} \geqslant 0$$

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

$$\underset{i\in I}{norm} = \frac{max\{x_i, 0\}}{\sum\limits_{i\in I} max\{x_i, 0\}}$$

Теорема 1. Пусть функция $R(\Phi, \Theta)$ непрерывно дифференцируема. Тогда точка локального экстремума (Φ^*, Θ^*) функции $Q(\Phi, \Theta)$ удовлетворяет системе уравнений

$$p_{twcd} = \underset{t\in T}{norm}\varphi_{wt}\theta_{ctd}$$

$$\varphi_{wt} = \underset{w\in W}{norm}\left(n_{wt} + \varphi_{wt}\frac{\partial R}{\partial \varphi_{wt}}\right) \qquad n_{wt} = \sum_{d\in D}\sum_{c\in C} n_{dwc}p_{twcd}$$

$$\theta_{ctd} = \underset{c\in C, t\in T}{norm}\left(n_{ctd} + \theta_{ctd}\frac{\partial R}{\partial \theta_{ctd}}\right) \qquad n_{ctd} = \sum_{w\in W} n_{dwc}p_{twcd}$$

где n_{dwc} частота отнесения термина w в документе d к классу c.

Доказательство:

Запишем Лагранжиан задачи при ограничениях нормировки, без учёта неотрицательности.

$$L(\Phi,\Theta)=\sum_{d\in D}\sum_{w\in W}\sum_{c\in C}n_{dwc}\log(\sum_{t\in T}\varphi_{wt}\theta_{ctd})+R(\Phi,\Theta)-\sum_{t\in T}\lambda_t\left(\sum_{w\in W}\varphi_{wt}-1\right)-\sum_{d\in D}\lambda_d\left(\sum_{c\in C}\sum_{t\in T}\theta_{ctd}-1\right)$$

Запишем необходимые условия локального экстремума задачи, по теореме Куна-Такера

$$\lambda_t=\sum_{d\in D}\sum_{c\in C}n_{dwc}\frac{\theta_{ctd}}{\sum\limits_{t\in T}\varphi_{wt}\theta_{ctd}}+\frac{\partial R}{\partial\varphi_{wt}}\sum_{t\in T}\varphi_{wt}\theta_{ctd}$$

Домножим обе части на φ_{wt} и просуммируем по всем терминам w.

$$\lambda_t=\sum_{w\in W}\left(n_{wt}+\varphi_{wt}\frac{\partial R}{\partial\varphi_{wt}}\right)$$

Снова домножим на φ_{wt}, подставим значения λ_t и выразим φ_{wt} и учтём ограничение неотрицательности, сделав положительную срезку $(x)_+=max(x,0)$.

$$\varphi_{wt}=\frac{\left(n_{wt}+\varphi_{wt}\frac{\partial R}{\partial\varphi_{wt}}\right)_+}{\left(\sum\limits_{w\in W}n_wt+\varphi_{wt}\frac{\partial R}{\partial\varphi_{wt}}\right)_+}=\underset{w\in W}{norm}\left(n_{wt}+\varphi_{wt}\frac{\partial R}{\partial\varphi_{wt}}\right)$$

Для θ_{ctd} аналогично получаем

$$\theta_{ctd}=\frac{\left(n_{ctd}+\theta_{ctd}\frac{\partial R}{\partial\theta_{ctd}}\right)_+}{\left(\sum\limits_{c\in C}\sum\limits_{t\in T}n_ctd+\theta_{ctd}\frac{\partial R}{\partial\theta_{ctd}}\right)_+}=\underset{c\in C,t\in T}{norm}\left(n_{ctd}+\theta_{ctd}\frac{\partial R}{\partial\theta_{ctd}}\right)$$

Регуляризация задачи

В этом разделе подберём на основании вероятностных предположений проблемно-ориентированные регуляризаторы. Обычно процесс регуляризации сводится к минимизации различия заданного распределения с модельным или к минимизации их взаимной ковариации. В роли функции различия двух распределений будет использовать дивергенцию Кульбака-Лейблера, характеризующую относительную энтропию двух различных дискретных распределений $(p_i)_{i=1}^n$ и $(q_i)_{i=1}^n$:

$$KL(p\parallel q)\equiv\sum_{i=1}^{n}p_i\ln\frac{p_i}{q_i}.$$

Минимизация KL-дивергенции эквивалента максимизации правдоподобия модельного распределения q по заданному p.

Разреживающий регуляризатор. Из предположения, что каждый документ и каждый термин связан с небольшим числом тем следует что для повышения качества модели нужно потребовать от матриц Φ и Θ сильную разреженность. Известно, что чем

меньше энтропия, тем более разрежено распределение. Поэтому будем максимизировать KL-дивергенция между заданным и равномерным для получения разреженного решения:

$$R(\Phi,\Theta) = -\beta_0 \sum_{t\in T}\sum_{w\in W} \beta_w \ln \varphi_{wt} - \alpha_0 \sum_{d\in D}\sum_{t\in T} \alpha_t \ln \theta_{ctd},$$

где $\alpha_t \in U(|T|), \beta_w \in U(|W|)$.

Декоррелирующий регуляризатор. С повышением различности тем повышается качество интерпретируемости модели. Для этой цели уменьшим ковариацию между столбцами φ_t и φ_s, где $t, s \in T$

$$R(\Phi) = -\gamma \sum_{t\in T}\sum_{s\in T\setminus t}\sum_{w\in W} \varphi_{wt}\varphi_{ws}$$

Сглаживающий регуляризатор. Минимизируем KL-дивергенцию между распределениями φ_t, θ_d и модельными β_w, α_w соответственно и воспользуемся аддитивностью регуляризаторов:

$$R(\Phi,\Theta) = \delta_0 \sum_{t\in T}\sum_{w\in W} \delta_w \ln \varphi_{wt} + \sum_{d\in D}\sum_{t\in T}\sum_{c\in C} \sigma_c\sigma_{ct} ln\theta_{ctd}$$

Отметим, что для векторов β_w, α_w взятых из распределения Дирихле получаются в точности формулы для М-шага модели LDA.

Разобьём множество тем на два подмножества, $T = S \cup B$: предметные темы S и фоновые B.

Предметные темы $t \in S$ содержат термины предметных областей. Их распределения $p(w|t)$ разрежены и декоррелированы. В силу того, что каждая предметная тема охватывает небольшое количество документов, распределения $p(c,t|d)$ тоже разрежены. *Фоновые темы* $t \in B$ содержат слова общей лексики. Их распределения $p(w|t)$, $p(c,t|d)$ сглажены, так как слова из этих тем содержаться во многих документах.

Воспользовавшись доказанной теоремой и аддитивностью регуляризаторов получим формулы для М-шага для комбинации описанных регуляризаторов:

$$\varphi_{wt} = \underset{w\in W}{norm}\left(n_{wt} - \beta_0\beta_w[t\in S] + \delta_0\delta_w[t\in B] - \gamma[t\in S]\varphi_{wt}\sum_{s\in T\setminus t}\varphi_{ws}\right)$$

$$\theta_{ctd} = \underset{c\in,t\in T}{norm}\left(n_{ctd} - \alpha_0\alpha_t[t\in S] + \sum_{c\in C}\sigma_c\sigma_{ct}[t\in B]\right)$$

EM - алгоритм

Для решение задачи будем применять итерационный процесс, состоящий из двух шагов - E(Expectation) и M(Maximization). Для первой итерацией выбирается начальное приближение параметров модели. На E-шаге пересчитываются частотные характеристики n_{dwc}, на M-шаге происходит обновление параметров модели φ_{wt}, θ_{ctd}.

Существуют модификации EM-алгоритма[**?**].

Обобщённый EM-алгоритм предлагает решать задачу M-шага не сверхточно, а лишь немного сдвигаясь к точке максимума правдоподобия, что приводит к более частому обновлению параметром φ_{wt} и θ_{ctd}. Данный подход увеличивает скорость сходимости алгоритма.

Тем не менее на больших коллекциях документов Обобщённый ЕМ-алгоритм может сходиться очень медленно. Например, в следствии разной скорости сходимости параметров модели. С помощью реорганизации порядка операций возможно добиться большей скорость сходимости. Этот подход используется в двух модификациях: *потоковом ЕМ-алгоритме* и *онлайновом ЕМ-алгоритме*. Последний из которых применим для динамически пополняющихся коллекций.

Статистический ЕМ-алгоритм

ЕМ-алгоритм крайне чувствителен к начальному приближению. Начальное приближение θ_{ctd}, φ_{wt} параметров можно задать, используя нормированные вектора из равномерного распределения и распределения Пуассона соответственно. Если же начальное приближение содержит нулевые элементы, то его рекомендуется смешать с неразреженным распределением.

Приведём ЕМ-алгоритм для построенной модели, ориентированный под языки программирования, использующие матричное умножение.

Алгоритм 1 ЕМ-алгоритм для задачи тематического моделирования

Вход: Коллекция документов D, число тем $|T|$, число классов $|C|$, начальное приближение Φ и Θ

Выход: Распределения Φ и Θ

пока Φ *и* Θ *не сойдутся* **выполнять**
  //Е-шаг
  обнулить n_{wt}, n_{ctd} для всех $d \in D, w \in W, c \in C, t \in T$
  для всех $d \in D, w \in W, c \in C$
    $Z_{dwc} = \sum\limits_{t \in T} \varphi_{wt}\theta_{ctd}$;
  конец цикла
  для всех $d \in D, w \in W, c \in C$
    $n_{dwc} = n_{dwc}/Z_{dwc}$;
  конец цикла
  для всех $w \in W, t \in T$
    $n_{wt} = \varphi_{wt}\left(\sum\limits_{d \in D}\sum\limits_{c \in C} n_{dwc}\theta_{ctd} + \frac{\partial R}{\partial \varphi_{wt}}\right)$;
  конец цикла
  для всех $c \in C, t \in T, d \in D$
    $n_{ctd} = \theta_{ctd}\left(\sum\limits_{w \in W} n_{dwc}\varphi_{wt} + \frac{\partial R}{\partial \theta_{ctd}}\right)$;
  конец цикла
  //М-шаг
  $n = \sum\limits_{d \in D}\sum\limits_{w \in W}\sum\limits_{c \in C} n_{dwc}\varphi_{wt}\theta_{ctd}$;
  $n_d = \sum\limits_{w \in W}\sum\limits_{c \in C} n_{dwc}$;
  $\varphi_{wt} = n_{wt}/n$;
  $\theta_{ctd} = n_{ctd}/n_d$;
конец цикла

Вычислительный эксперимент

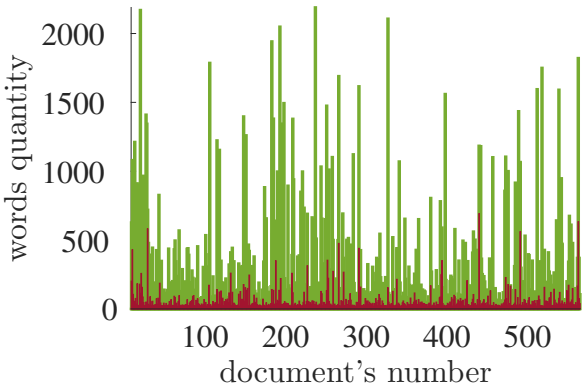
В этом разделе используем построенную тематическую модель для создания рекомендательной системы. Будем использовать данные мобильное приложения для изучения иностранных слов. Будем считать, что нужно предлагать пользователю слова на момент сформирования выборки.

Принцип работы приложения. Мобильное приложение предлагает случайным образом пользователю слова одно за другим из словаря. Каждое из предложенных слов пользователь относит к одному из двух классов: принятые(added) и отброшенные(refused). Слова из класса *added* пользователь будет запоминать в дальнейшем при помощи приложения. Слова из класса *refused* пользователь по каким-то причинам не желает запоминать(то ли он их уже знает, то ли они ему не интересны). Каждый день пользователю может выбрать не более 10 новых слов для запоминания.

Исходные данные. Для каждого из классов изначальная выборка состояла из пар (пользователь, слово). Естественно считать, что если пользователь отказался от изучения слова, то в дальнейшем оно ему не предлагалась. Исходя из этого, у каждого из пользователей были удалены те слова из класса added, которые имелись в классе refused.

Таблица 1. Характеристики выборки

	added	refused
число слов:	163674	33024
максимальное число слов у пользователя:	2221	692
минимальное число слов у пользователя:	20	20
количество пользователей:	557	557
размер словаря:	15403	15403



(a) Зелёный: added. Красный: refused.

Переформулируем задачу в терминах тематических моделей. С каждым пользователем связана история предложенных ему слов и его предпочтения (релевантно/нет) относительно них, подобно тому как с документом связаны термины, из которых он состоит. Поэтому будем рассматривать пользователя как документ d, а предложенные ему слова и его предпочтения – как термины w и классы c соответственно. В данной задаче клас-

130 сов два (+ и -). Будем считать, что не важен порядок выбранных и отброшенных слов и
131 порядок пользователей. Положим распределение слов по темам общим для всей выбор-
132 ки и не зависящим от пользователя. Предположим, что каждый пользователь обработает
133 небольшое число слов из коллекции и к каждому классу относится небольшое число слов.
134 Базируясь на этих предположениях построим тематическую модель для классификации
135 предлагаемых слов для каждого из пользователей.

Рассмотрим вид матрицы частот $N = (n_{dwc})_{D\times W\times C}$. Она состоит из нулей и единиц. Естественно предположить, что выборка распределена по закону Бернули. Тогда нахождение параметров модели Φ, Θ сведётся к максимизации следующего функционала:

$$Q^*(\Phi,\Theta) = \sum_{d\in D}\sum_{w\in W}\sum_{c\in C}\sum_{n_{dwc}=1} n_{dwc}\log(\sum_{t\in T}\varphi_{wt}\theta_{ctd}) +$$

$$\sum_{d\in D}\sum_{w\in W}\sum_{c\in C}\sum_{n_{dwc}=0}(1-n_{dwc})\log(1-\sum_{t\in T}\varphi_{wt}\theta_{ctd}) + R(\Phi,\Theta)$$

Теорема 2. Пусть функция $R(\Phi,\Theta)$ непрерывно дифференцируема, $n_{dwc}\in\{0,1\}$ Тогда точка локального экстремума (Φ^*,Θ^*) функции $Q^*(\Phi,\Theta)$ удовлетворяет системе уравнений

$$p_{twcd}(n_{dwc}=1) = \underset{t\in T, n_{dwc}=1}{norm}\varphi_{wt}\theta_{ctd} \qquad p_{twcd}(n_{dwc}=0) = \frac{-\varphi_{wt}\theta_{ctd}}{1-\sum\limits_{t\in T}\sum\limits_{n_{dwc}=0}\varphi_{wt}\theta_{ctd}}$$

$$\varphi_{wt} = \underset{w\in W}{norm}\left(n_{wt} + \varphi_{wt}\frac{\partial R}{\partial\varphi_{wt}}\right) \qquad n_{wt} = \sum_{d\in D}\sum_{c\in C} p_{twcd}$$

$$\theta_{ctd} = \underset{c\in C, t\in T}{norm}\left(n_{ctd} + \theta_{ctd}\frac{\partial R}{\partial\theta_{ctd}}\right) \qquad n_{ctd} = \sum_{w\in W} p_{twcd}$$

136 **Измерение качества модели.**

Точность тематической модели $p(c,w|d)$ на коллекции текстовых документов D принято измерять с помощью *перплексии*, которая является преобразованием от функции правдоподобия:

$$P(D,p) = exp\left(-\frac{1}{n}L(\Phi,\Theta)\right)$$

Вероятность отнесения термина w в документе d к классу равна:

$$p(c|w,d) = \frac{p(c,w|d)}{p(w|d)} = \frac{p(c,w|d)}{\sum\limits_{c\in C}p(c,w|d)} = \frac{\sum\limits_{t\in T}\varphi_{wt}\theta_{ctd}}{\sum\limits_{c\in C}\sum\limits_{t\in T}\varphi_{wt}\theta_{ctd}}$$

137 Выборку $X = (x_i, c_i)_{i=1}^l$ разделим на две части: обучающую $X^m = (x_i, c_i)_{i=1}^m$ и контрольную
138 $X^l = (x_i, c_i)_{i=1}^l$, где $x_i = (w_i, d_i)$. Введём функцию весов следующим образом:

$$w_c(\nu_A) = \begin{cases} \frac{1}{\exp(\nu_A - 1)} & \text{для с =+,} \\ \frac{1}{1-\exp(\nu_A-1)} & \text{для с =-,} \end{cases}$$

139 где $\nu_A = \dfrac{\sum\limits_{w\in W} n_{dwc}}{\sum\limits_{cinC}\sum\limits_{w\in W} n_{dwc}}$

Тогда алгоритм классификации имеет следующий вид:

$$\mathsf{a}(x, X^l) = \arg\underset{c\in C}{max}\; p(c|x)\mathrm{w}_c(\nu_A),$$

Качество классификации для каждого из классов в отдельности общепринято оценивать по *точности(precision)*, *полноте(recall)* и *F1-мере*.

$$precision(c^*) = \frac{\sum\limits_{i=1}^{l}[a(x_i) = c_i(x_i) \wedge c_i(x_i) = c^*]}{\sum\limits_{i=1}^{l}[a(x_i) = c^*]}$$

$$recall(c^*) = \frac{\sum\limits_{i=1}^{l}[a(x_i) = c_i(x_i) \wedge c_i(x_i) = c^*]}{\sum\limits_{i=1}^{n}[c_i(x_i) = c^*]}$$

$$F1(c^*) = \frac{precision(c^*) * recall(c^*)}{precision(c^*) + recall(c^*)}$$

Для оценки качества классификации всех классов используются критерии F1(macro) и F1(micro). Последний из них в случае когда классов два в точности совпадает с accuracy.

$$F1(macro) = \frac{1}{|C|}\sum_{c \in C} F1(c)$$

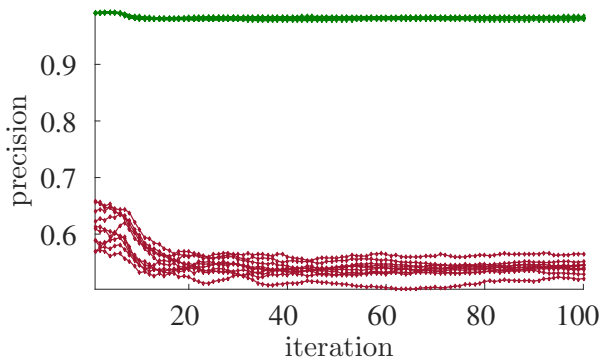
$$accuracy = \frac{1}{n}\sum_{i=1}^{l}[a(x_i) = c_i(x_i)]$$

Результаты эксперимента. В эксперименте было выбрано фиксированное число тем $|T| = 50$, из них 30 тем были положены *предметными* $|S| = 30$, оставшиеся 20 тем – *фоновыми* $|B| = 20$. Число итераций было выбрано равным 100, такое число обусловлено малым изменением норм матриц Φ, Θ. За начальное приближение матриц Φ, Θ были взяты вектора из распределений Пуассона и равномерного соответственно.

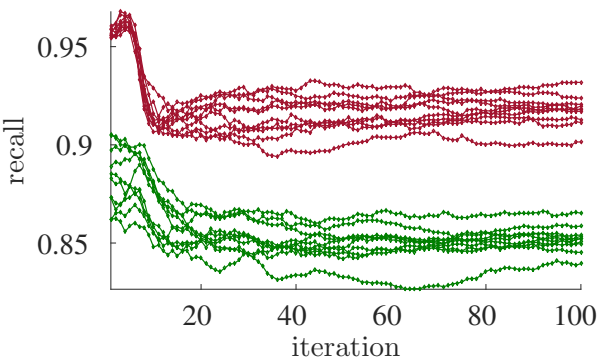
Для разреживания предметных тем были взяты векторы из равномерного распределения, с коэффициентами регуляризации $\beta_0 = 1$, $\delta_0 = 10^{-3}$ и $\alpha_0 = 1$. Для сглаживание фоновых тем были выбраны векторы из Бета-распределения с параметрами, с коэффициентами регуляризации $\sigma_{ct} = 1\ \forall c \in C$. Коэффициент при декоррелирующем регуляризаторе $\gamma = 5 * 10^3$. Значения коэффициентов на протяжении всех итерации оставались неизменными. Декорялирующий и сглаживающий регуляризатор включался в модель начиная с первой итерации. Разреживание предметных тем было запущено только после того, как итерационный процесс начал сходиться.

Кросс-валидация. Исходная выборка была разделена на 10 частей случайным образом для каждого из пользователя и каждого класса. Каждая из частей выступала в роли контрольной выборки, по оставшимся 9 частям происходило обучение модели.

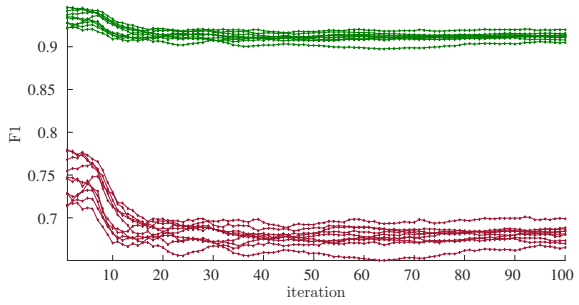
Процесс построения модели контролировался по графикам качества модели от итерации. По ним же были подобраны коэффициенты регуляризации. Разреживание предметных тем рекомендуется включать после того, как итерационный процесс начнёт сходиться. Как можно видеть из графика для перплексии, в поставленном эксперименте сходимость наблюдается после 20-ти итерации. Декоррелирование предметных тем и сглаживание фоновых были включены, начиная с первой итерации.



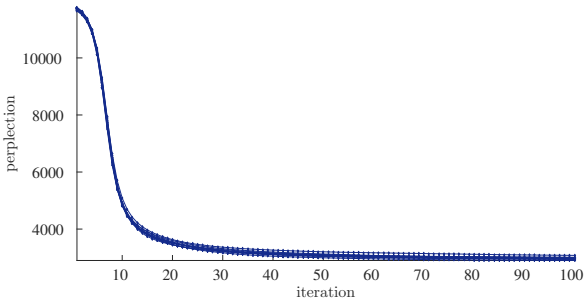
(b) Зелёный: added. Красный: refused.



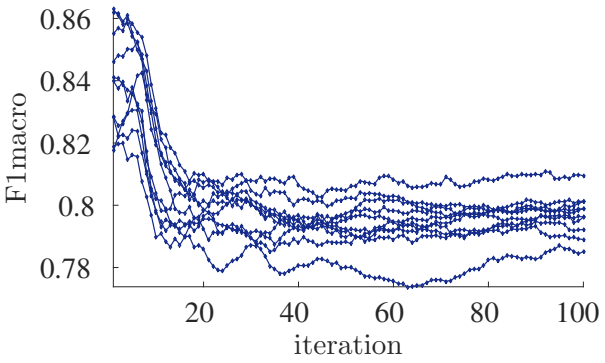
(c) Зелёный: added. Красный: refused.



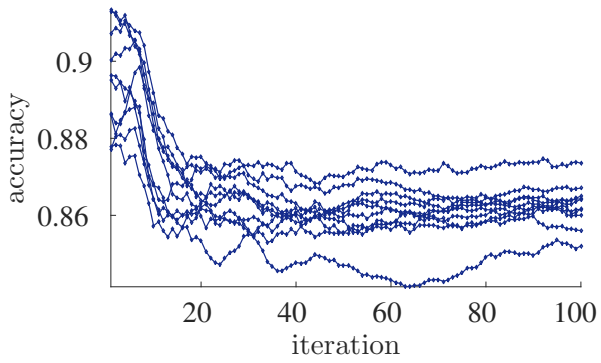
(d) Зелёный: added. Красный: refused.



(e) Perplection



(f) F1(macro)



(g) Accuracy

Таблица 2. Критерии качества

	F1	precision	recall
Added:	0.9124	0.9819	0.8520
Refused:	0.6813	0.5420	0.9175

Таблица 3. Caption

F1(macro)	accuracy
0.7968	0.8625

Заключение

В данной работе предложен метод построения тематической модели для задачи классификации, основанный на подходе аддитивной регуляризации тематических моделей и осуществлён подбор предметно-ориентированных регуляризаторов. Предложен EM-алгоритм для решения полученной оптимизационной задачи. Предложена модификация полученной модели для построения рекомендательной системы. Проведён вычислительный эксперимент на реальных данных и вычислены критерии качества рекомендаций.

Литература

[1] Michael I. Jordan David M. Blei, Andrew Y. Ng. Latent dirichlet allocation. *Journal of Machine Learning Research*, 2003.

[2] Potapenko A. A. Vorontsov K. V. Additive regularization of topic models. *Machine Learning. Special Issue "Data Analysis and Intelligent Optimization with Applications*, 2014.

[3] Воронцов К.В. Вероятностное тематическое моделирование. *http://www.machinelearning.ru/wiki/images/2/22/Voron-2013-ptm.pdf*, 2013.

[4] Thomas Hoffman. Probabilistic latent semantic indexing. *Proceedings of the Twenty-Second Annual International SIGIR Conference on Research and Development in Information Retrieval*, 1999.