

МИНОБРНАУКИ РОССИИ
Федеральное государственное бюджетное образовательное учреждение
высшего образования

**«САРАТОВСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
ИМЕНИ Н. Г. ЧЕРНЫШЕВСКОГО»**

Кафедра математической кибернетики и компьютерных наук

**СОЗДАНИЕ ФРЕЙМВОРКА ДЛЯ РАБОТЫ С
АВТОМАТИЗИРОВАННЫМ ТЕСТИРОВАНИЕМ**

КУРСОВАЯ РАБОТА

Студента 3 курса 351 группы
направления 09.03.04 — Программная инженерия
факультета КНиИТ
Кондрашова Даниила Владиславовича

Научный руководитель
зав.каф.техн.пр.,
доцент, к. ф.-м. н.

С. В. Папшев

Заведующий кафедрой
к. ф.-м. н.

С. В. Миронов

Саратов 2024

СОДЕРЖАНИЕ

ВВЕДЕНИЕ	3
1 Математические основы тематического моделирования	4
1.1 Основная гипотеза тематического моделирования	4
1.2 Аксиоматика тематического моделирования	4
1.3 Задача тематического моделирования	5
1.4 Метод решения обратной задачи	5
1.4.1 Сведение обратной задачи к задаче максимизации функ- ционала	6
1.5 название не придумалл	7
1.6 Сглаживание	8
1.7 Разреживание	8
1.8 Декоррелирование	8
ЗАКЛЮЧЕНИЕ	8
СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ	8

ВВЕДЕНИЕ

С ростом объёмов информации в наше время для её эффективного поиска и изучения стало необходимо уметь её классифицировать и структурировать. Сегодня физически невозможно найти нужные данные просто перебирая все ресурсы подряд, появилась острая потребность в поиске по темам, в классификации данных.

Данную проблему призвано решить тематическое моделирование. Оно способно быстро и эффективно автоматически разбить по темам огромные объёмы информации.

1 Математические основы тематического моделирования

1.1 Основная гипотеза тематического моделирования

Тематическое моделирование — это подход анализа текстовых данных, направленный на выявление семантических структур в коллекции документов.

Само тематическое моделирование основывается на предположении, что слова в тексте связаны не с документом, а с темой. Кроме того первично текст разбивается на темы, затем каждая из них порождает слово для соответствующих позиций в документе. Таким образом, сначала порождается тема, а потом термины.

Благодаря этой гипотезе можно по частоте и встречаемости слов производить тематическую классификацию текстов.

1.2 Аксиоматика тематического моделирования

Каждый текст можно количественно охарактеризовать. Вот основные количественные характеристики, используемые при тематическом моделировании:

- W — конечное множество термов;
- D — конечное множество текстовых документов;
- T — конечное множество тем;
- $D \times W \times T$ — дискретное вероятностное пространство;
- коллекция — i.i.d выборка $(d_i, w_i, t_i)_{i=1}^n$;
- $n_{dwt} = \sum_{i=1}^n [d_i = d][w_i = w][t_i = t]$ — частота (d, w, t) в коллекции;
- $n_{wt} = \sum_d n_{dwt}$ — частота термина w в документе d ;
- $n_{td} = \sum_w n_{dwt}$ — частота термов темы t в документе d ;
- $n_t = \sum_{d,w} n_{dwt}$ — частота термов темы t в коллекции;
- $n_{dw} = \sum_t n_{dwt}$ — частота термина w в документе d ;
- $n_W = \sum_d n_{dw}$ — частота термина w в коллекции;
- $n_d = \sum_w n_{dw}$ — длина документа d ;
- $n = \sum_{d,w} n_{dw}$ — длина коллекции.

Также в тематическом моделировании используются следующие гипотезы и аксиомы:

- Независимость слов от порядка в документе: порядок слов в документе не важен;
- Независимость от порядка документов в коллекции: порядок документов

в коллекции не важен;

- Зависимость термина от темы: каждый терм связан с соответствующей темой и порождается ей;
- Гипотеза условной независимости: $p(w|d, t) = p(w|t)$.

Вышеперечисленные характеристики, гипотезы и аксиомы являются основой тематического моделирования, являющейся достаточной для построения тематической модели.

1.3 Задача тематического моделирования

Как уже говорилось ранее, документ порождается следующим образом:

1. для каждой позиции в документе генерируется тема $p(t|d)$;
2. для каждой сгенерированной темы в соответствующей позиции генерируем терм $p(w|d, t)$.

Тогда вероятность появления слова в документе можно описать по формуле полной вероятности:

$$p(w|d) = \sum_{t \in T} p(w|d, t)p(t|d) = \sum_{t \in T} p(w|t)p(t|d) \quad (1)$$

Такой алгоритм является прямой задачей порождения текста. Тематическое моделирование призвано решить обратную задачу:

1. для каждого термина w в тексте найти вероятность появления в теме t (найти $p(w|t) = \phi_{wt}$);
2. для каждой темы t найти вероятность появления в документе d (найти $p(t|d) = \theta_{td}$).

Обратную задачу можно представить в виде стохастического матричного разложения.

Таким образом, тематическое моделирование ищет величину $p(w|d)$.

1.4 Метод решения обратной задачи

Для решения задачи тематического моделирования необходимо найти величину $p(w|d)$, сделать это можно с помощью метода максимального правдоподобия.

1.4.1 Сведение обратной задачи к задаче максимизации функционала

$$\prod_{i=1}^n p(d_i, w_i) = \prod_{d \in D} \prod_{w \in d} p(d, w)^{n_{dw}}, \quad (2)$$

следовательно,

$$\sum_{d \in D} \sum_{w \in d} n_{dw} \ln p(w|d) \xrightarrow{const} n_{dw} \rightarrow \max \quad (3)$$

Данная задача эквивалентна задаче максимизации функционала

$$L(\Phi, \Theta) = \sum_{d \in D} \sum_{w \in d} n_{dw} \ln \sum_{t \in T} \phi_{wt} \theta_{td} \rightarrow \max_{\Phi, \Theta} \quad (4)$$

при ограничениях неотрицательности и нормировки

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

Из ограничений следует, что сумма элементов столбцов матриц равна единице и каждый их элемент больше или равен нулю, следовательно,

Пусть $\Omega = (w_j)_{j \in J}$ - набор нормированных неотрицательных векторов $w_j = (w_{ij})_{i \in I_j}$ различных размерностей $|I_j|$.

Задача максимизации функции $f(\Omega)$ на единичных симплексах:

$$\begin{cases} f(\Omega) \rightarrow \max_{\Omega}; \\ \sum_{i \in I_j} w_{ij} = 1, \quad j \in J; \\ w_{ij} \geq 0, \quad i \in I_j, \quad j \in J. \end{cases} \quad (5)$$

Операция нормировки вектора: $p_i = (x_i)_{i \in I} = \frac{\max x_i, 0}{\sum_{k \in I} \max x_k, 0}$.

Лемма о максимизации функции на единичных симплексах:

Пусть $f(\Omega)$ непрерывно дифференцируема по Ω . Тогда векторы w_j локаль-

ного экстремума задачи $f(\Omega) \rightarrow \max$ удовлетворяют системе уравнений:

$$\begin{cases} w_{ij} = \underset{i \in I_j}{\text{norm}} \left(w_{ij} \frac{\partial f}{\partial w_{ij}} \right), & \text{если } \exists i : w_{ij} \frac{\partial f}{\partial w_{ij}} > 0 \\ w_{ij} = \underset{i \in I_j}{\text{norm}} \left(-w_{ij} \frac{\partial f}{\partial w_{ij}} \right), & \text{иначе, если } \exists i : w_{ij} \frac{\partial f}{\partial w_{ij}} < 0 \\ w_{ij} \frac{\partial f}{\partial w_{ij}} = 0, & \text{иначе} \end{cases} \quad (6)$$

Тогда исходная задача принимает следующий вид:

$$\begin{cases} p_{tdw} = p(t|d, w) = \underset{t \in T}{\text{norm}}(\phi_{wt}\theta_{td}) \\ \phi_{wt} = \underset{w \in W}{\text{norm}}(n_{wt}) \\ \theta_{td} = \underset{t \in T}{\text{norm}}(n_{td}) \end{cases} \quad (7)$$

1.5 название не придумалл

Описанная в прошлом разделе модель не является корректно заданной, так как у этой системы уравнений может быть несколько решений, следовательно решение нужно конкретизировать с помощью регуляризации.

Регуляризация — стандартный приём доопределения решения с помощью дополнительных критериев.

Таким образом, добавим некоторый регуляризатор $R(\Phi, \Theta) = \sum_i \tau_i R_i(\Phi, \Theta)$ к задаче максимизации логарифма правдоподобия:

$$\sum_{d,w} n_{dw} \ln \sum_{t \in T} \phi_{wt} \theta_{td} + R(\Phi, \Theta) \rightarrow \max_{\Phi, \Theta}; \quad (8)$$

Тогда система уравнения примет вид:

$$\begin{cases} p_{tdw} = p(t|d, w) = \underset{t \in T}{\text{norm}}(\phi_{wt}\theta_{td}) \\ \phi_{wt} = \underset{w \in W}{\text{norm}} \left(n_{wt} + \phi_{wt} \frac{\partial R}{\partial \phi_{wt}} \right) \\ \theta_{td} = \underset{t \in T}{\text{norm}} \left(n_{td} + \theta_{td} \frac{\partial R}{\partial \theta_{td}} \right) \end{cases} \quad (9)$$

1.6 Сглаживание

1.7 Разреживание

1.8 Декоррелирование

ЗАКЛЮЧЕНИЕ

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ