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

В настоящее время рост информации привёл к проблеме трудности её обработки. Информацию необходимо перед изучением как-то классифицировать, структурировать и систематизировать. Инструмент, который способен помочь в решении данной задачи — это тематическое моделирование.

Чаще всего тематическое моделирование не является самоцелью, однако оно необходимо для решения многих других задач, таких как разведочный поиск, и так далее.

Трудно переоценить всю важность тематического моделирования, если представить, поиск нужной информации не в тематическом разделе, а просто по всем документам.

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 зиждится 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 взаимовстречаемости слов производить тематическую классификацию текстов.

1.2 Аксиоматика тематического моделирования

Каждый текст можно количественно охарактеризовать. Вот основные количественные характеристики, использующиеся 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 терма 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 терма w в документе d;
— $n_W = \sum_d n_{dw}$ — частота терма 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 терма от темы: каждый терм связан с соответствующей темой и порождается ей;

— Гипотеза условной независимости: $p(w|t, d) = p(w|t)$.

Вышеперечисленные характеристи, 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:

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

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

1. для каждого терма 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 название не придумал

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

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

следовательно,

$$\sum_{d\in D}\sum_{w\in d} n_{dw}\ln p(w|d)\underset{const}{\cancel{p(d)}} = n_{dw} \to max \tag{3}$$

Данная задача эквивалентна задачи максимизации функционала

$$L(\Phi,\Theta) = \sum_{d\in D}\sum_{w\in d} n_{dw}\ln\sum_{t\in T}\phi_{wt}\theta_{td} \to \underset{\Phi,\Theta}{max} \tag{4}$$

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 равен больше или равен нулю. Тогда задачу можно свести к задаче максимизации функции на единичных симплексах:

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:

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

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

Лемма о максимизации функции на единичных симплексах:

Пусть $f(\Omega)$ непрерывно дифференцируема по Ω. Тогда векторы w_j локального экстремума задачи $f(\Omega) \to \max$ удовлетворяют системе уравнений:

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

Тогда исходная задача принимает следующий вид:

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

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:

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

Тогда система уравнения примет вид:

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

1.6 Сглаживание

1.7 Разреживание

1.8 Декоррелирование

ЗАКЛЮЧЕНИЕ

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ