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 Элементарное решение обратной задачи

Тематическое моделирование ищет величину $p(w|d)$, чтобы её вычислить нужно знать вероятности $p(w|t) = \phi_{wt}$ и $p(t|d) = \theta_{td}$.

По определению:

$$\phi_{wt} = \frac{n_{wt}}{n_t} = \frac{\sum_d n_{dwt}}{\sum_{d,w} n_{dwt}} \tag{2}$$

и

$$\theta_{td} = \frac{n_{td}}{n_d} = \frac{\sum_w n_{dwt}}{\sum_{t,w} n_{dwt}} \tag{3}$$

Следовательно, для решения задачи, осталось вычислить величину n_{dwt}. Выразим $p(t|d,w) = \frac{n_{dwt}}{n_{dw}}$ через матрицы ϕ_{wt}, θ_{td} по формуле Байеса:

$$p(t|d,w) = \frac{p(w,t|d)}{p(w|d)} = \frac{p(w|t)p(t|d)}{p(w|d)} = \frac{\phi_{wt}\theta_{td}}{\sum_s \phi_{ws}\theta_{ws}} \tag{4}$$

Туперь задачу тематического моделирования можно переписать в виде системы уравнений относительно $\phi_{wt}, \theta_{td}, n_{dwt}$:

$$\begin{cases} n_{dwt} = n_{dw}\frac{\phi_{wt}\theta_{td}}{\sum_s \phi_{wt}\theta_{td}}, & d \in D, w \in W, t \in T \\ \phi_{wt} = \frac{\sum_d n_{dwt}}{\sum_{d,w} n_{dwt}}, & w \in W, t \in T \\ \theta_{td} = \frac{\sum_w n_{dwt}}{\sum t,w n_{dwt}}, & d \in D, t \in T \end{cases} \tag{5}$$

ЗАКЛЮЧЕНИЕ

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ