ВИКОРИСТАННЯ АДИТИВНОЇ РЕГУЛЯРИЗАЦІЇ ПРИ РОЗВ'ЯЗАННІ ЗАДАЧІ ТЕМАТИЧНОГО ОЦІНЮВАННЯ

Деркач О. С.
Науковий керівник – к.т.н., доц. Гибкіна Н.В.
Харківський національний університет радіоелектроніки
(61166, Харків, пр. Науки, 14, каф. Прикладної математики,
тел. (057) 702-14-36)
e-mail: oleksii.derkach@nure.ua

Topic modeling is one of the modern directions of the statistic processing of natural language, which has been actively developing since the late 1990s. The topic modeling method allows to build a model, which determines the topics probability distribution for each document. Different regularizes can be applied during learning process to give the special properties for the solution.

У зв'язку зі зростанням інформаційних потоків, інтенсивним накопиченням інформації, розвитком масової та ділової комунікації, розповсюдженням соціальних мереж та інших інтернет-ресурсів актуальними стають задачі вилучення інформації та її аналізу. Цей аналіз дозволяє збирати дані про тематику, настрій, авторство тексту, що в подальшому спрощує автоматизовану роботу з вихідними текстами. Для розв'язання подібних задач перспективними вважаються різноманітні математичні методи, оскільки вони дозволяють робити науково обґрунтовані висновки. Найбільш популярними математичними методами вилучення інформації з тексту є статистичні методи обробки текстів. До них відноситься метод тематичного моделювання, що дозволяє будувати модель колекції документів, яка визначає тематичне направлення кожного з них.

Метою тематичного моделювання є визначення тематики документів та пов'язаних з ними об'єктів.

Перед будуванням тематичних моделей текст природньої мови зазвичай підлягає серії перетворень: лематизації, стемінгу, видаленню стоп-слів та рідкісних слів. Багато з цих перетворень здійснюються за допомогою алгоритмів, реалізованих у бібліотеках різних мов програмування, отже, їх будемо використовувати для попередньої обробки тексту.

Позначимо D – множина (колекція) текстових документів, W – множина (словник) всіх термів, які вживаються [1]. Термами можуть бути нормальні форми слів, словосполучення або терміни. Кожний документ $d \in D$ являє собою послідовність слів $w_1,...w_{n_d} \in W$, де n_d – довжина документа d в термах. Кожне входження терму w у документ d пов'язано з деякою темою t зі скінченної множини T.

Відповідно до формули повної ймовірності та гіпотези умовної незалежності розподіл термів в документі $p(w|d)$ описується ймовірнісною сумішшю розподілів термів в темах $\varphi_{wt} = p(w|d)$ з вагами $\theta_{td} = p(t|d)$:

$$p(w|d) = \sum_{t\in T} p(w|t,d)p(t|d) = \sum_{t\in T} p(w|t)p(t|d) = \sum_{t\in T} \varphi_{wt}\theta_{td} \,. \qquad (1)$$

Задача тематичного моделювання є оберненою задачею: по заданій колекції D необхідно знайти параметри φ_{wt}, θ_{td}, за яких тематична модель добре наближає частотні оцінки умовних ймовірностей.

Рівність (1) можна переписати у матричному вигляді $F \approx \Phi\Theta$, де $F = (p(w|d))_{W\times D}$ – відома матриця частот термів в документах $F = (p(w|d))_{W\times D}$, а $\Phi = (\varphi_{wt})_{W\times T}$ та $\Theta = (\theta_{td})_{T\times D}$ – невідомі матриці термів тем та тем документів відповідно. Тоді задача тематичного моделювання зводиться до пошуку наближеного матричного розкладання $F \approx \Phi\Theta$. Зауважимо, що всі три матриці є стохастичними, тобто мають невід'ємні нормовані стовпці.

Для знаходження параметрів моделі зручно використовувати ЕМ-алгоритм, що складається з двох кроків. На першому кроці знаходять оцінки апостеріорного розподілу $p(t|d,w)$ за вибраних параметрів моделі φ_{wt} та θ_{td}. На другому кроці алгоритму потрібно оновити параметри моделі. Це можна зробити максимізуючи математичне сподівання логарифму правдоподібності апостеріорного розподілу. В результаті отримуємо ітеративний алгоритм, який розпочинається з деякого наближення φ_{wt}^0 та θ_{td}^0.

Задача стохастичного матричного розкладання є некоректно поставленою, оскільки множина її розв'язків у загальному випадку нескінченна. Існує загальний підхід до розв'язання некоректно поставлених обернених задач, який називається регуляризацією. Регуляризатор враховує специфіку розв'язуваної задачі та знання предметної області. Наприклад, він може надавати властивості розрідженості, або, навпаки, згладити ймовірності по темам.

Адитивна регуляризація тематичних моделей матиме вигляд:

$$\sum_{d\in D}\sum_{w\in d} n_{dw} \ln \sum_{t\in T} \varphi_{wt}\theta_{td} + \sum_{i=1}^{k} \tau_i R_i(\Phi,\Theta) \to \max_{\Phi,\Theta},$$

$$\sum_{w\in W} \varphi_{wt} = 1, \;\; \sum_{t\in T} \theta_{td} = 1, \;\; \varphi_{wt} \geq 0, \;\; \theta_{td} \geq 0,$$

де $R_i(\Phi,\Theta)$ – регуляризатори з коефіцієнтами регуляризації $\tau_i \geq 0$, $i = \overline{1,k}$.

В роботі порівнюються результати розв'язання задачі тематичного моделювання з використанням регуляризаторів згладжування та розрідження, декорелірування, визначення кількості тем.

Список використаних джерел:

1. Apishev M., Koltcov S., Koltsova O., Nikolenko S., Vorontsov K. (2017) Additive Regularization for Topic Modeling in Sociological Studies of User-Generated Texts. In: Sidorov G., Herrera-Alcántara O. (eds) Advances in Computational Intelligence. MICAI 2016. Lecture Notes in Computer Science, vol 10061. Springer, Cham. PP.169-184.