УДК 004.81
AMS MSC2020: 68T99

Метод максимального правдоподобия для обобщения нечетких множеств в таксономиях

Айрапетян Ж. С., Фролов Д. С., Миркин Б. Г.

НИУ «Высшая школа экономики»

Аннотация. В работе предлагается новый метод обобщения тематической текстовой коллекции, оснащенной таксономией предметной области. С помощью спектральных методов кластеризации из текстовой коллекции извлекаются нечеткие множества листьев таксономии, соответствующие понятиям, одновременно используемым в статьях коллекции. Эти нечеткие множества обобщаются путем их подъема в дереве таксономии с использованием критерия максимального правдоподобия. Оптимальный подъем подразумевает нахождение вершины или множества вершин в дереве таксономии, наиболее плотно покрывающих листовые понятия из обобщаемого множества. Наш метод включает два основных этапа: (1) извлечение кластеров из текстовой коллекции и (2) обобщение этих кластеров. В данной работе модернизируются оба этапа. Алгоритмы применены к структурному анализу и описанию текстовой коллекции из 17 тыс. аннотаций научных статей в области Наук о данных, опубликованных в журналах издательства Шпрингер. Таксономия Наук о данных, используемая в данной работе, является шестиуровневой иерархической таксономией, разработанной вручную международной Ассоциацией Вычислительной Техники и Вычислительных Систем (ACM-CSS [4])

Ключевые слова: иерархическая таксономия, методы обобщения, спектральная кластеризация, аннотированные суффиксные деревья.

Введение

Вопросы автоматизации анализа текстовых коллекций приобретают все большее значение, как в силу практических потребностей, так

и в силу теоретической необходимости. В работе исследуется математический аналог уникальной когнитивной способности человека — обобщения. Понятие обобщения в данном контексте подразумевает извлечение концепций большего объема, но менее конкретного содержания из коллекции документов, то есть перехода от частного к общему согласно разработанному нами подходу [3]. Публикации в области анализа текстовых коллекций указывают на иерархическую структуру концепций. Такая иерархическая структура строится переходами от общего к частному, а значит, напрямую затрагивает понятие обобщения. Таксономия предметной области представляет собой иерархическую структуру корневого дерева. Первым делом из текстовой коллекции извлекаются нечеткие множества листьев таксономии, отражающие структуру текстовой коллекции, а затем эти множества обобщаются с использованием структуры таксономии. Такая процедура позволяет выявить узлы таксономии, наиболее точно описывающие текстовую коллекцию.

1. Извлечение кластеров

Для извлечения нечетких кластеров на множестве листьев таксономии необходимо вычислить релевантность каждого текста к каждой листовой теме таксономии. Для вычисления матрицы релевантностей R используются Аннотированные Суффиксные деревья (AST [2]), построенные для каждого текста из коллекции. Матрица ко-релевантности листовых тем A вычисляется как $A = R^T R$, то есть для двух листовых тем i и j с векторами релевантности текстам r_i и r_j их схожесть определяется как скалярное произведение. Мы также используем веса текстов, учитывающие их уникальность.

Теперь задачу кластеризации можно формулировать в терминах оптимального разбиения взвешенного графа, представляемого матрицей A, на кластеры. Такую задачу можно решить с помощью комбинации двух алгоритмов: LaplacianEigenMaps [1] и метода нечетких C-средних. Матрица смежности A используется для вычисления матрицы нормализованного Лапласиана L_n , спектр которого (особенно собственные векторы, соответствующие наименьшим собственным значениям) имеет интерпретацию в терминах минимального числа разрезов, необходимого для разделения графа на компоненты сопо-

ставимого размера. Затем полученные вложения кластеризуются с помощью метода нечеткой кластеризации C-средних.

2. Методы обобщения

После извлечения нечетких кластеров из текстовой коллекции, мы хотим обобщить эти кластеры, используя дерево таксономии. Введем некоторые обозначения и определения для удобства дальнейших рассуждений. Пусть I — множество листьев нашей таксономии T, где T — множество всех узлов таксономии, тогда для любого $h \in T \setminus I$: $T(h)$ будет поддеревом дерева T, а $I(h)$ — терминальными узлами этого поддерева. Для любого $h \in T \setminus I$ будем обозначать множество непосредственных отпрысков узла h как $\chi(h)$. Дадим определение нечеткому множеству. Нечеткое множество на листьях I нашего дерева — это отображение $u : I \to \mathbb{R}^{[0,1]}$, где $u(i)$ для $i \in I$ — это значение принадлежности листового элемента данному множеству. $S_u = \{i \in I : u(i) > 0\}$ — основа нечеткого множества u. Все вершины $t \in T$, такие что $I(t) \cap S_u = \varnothing$ будем называть u-нерелевантными вершинами. Из такого определения вытекает, что для любой u-нерелевантной вершины ее потомки тоже являются u-нерелевантными. Максимально u-нерелевантным узлом называется узел, который является u-нерелевантным, а его родитель уже не является u-нерелевантным. Удалим из дерева таксономии все не максимальные u-нерелевантные узлы. Теперь все u-нерелевантные узлы — это листья, не входящие в основу нечеткого множества. В работе [3] была рассмотрена следующая задача обобщения: дано нечеткое множество на терминальных узлах таксономии T, требуется найти вершину $h \in T$, которая покрывала бы данное множество настолько плотно, насколько это возможно. Набор узлов H будем называть u-покрытием, если:

(а) H покрывает S_u, то есть $S_u \subseteq \bigcup_{h \in H} I(h)$;

(б) узлы в H не связаны, то есть $I(h) \cap I(h') = \varnothing$ для любых $h, h' \in H$ таких, что $h \neq h'$.

Узлы, принадлежащие $H \setminus I$, будут являться головными темами покрытия, а узлы, принадлежащие $H \cap I$, индуцированными ею

выбросами. Множеством пробелов покрытия будет являться объединение по всем индуцированным покрытием u-нерелевантым узлам, то есть $\bigcup\limits_{h\in H\setminus I} I(h)\setminus S_u$. Обозначим $G(h)=I(h)\setminus S_u$. Значение штрафной функции, ассоциированное с определенным u-покрытием должно учитывать принадлежности $u(i)$ листовых элементов своих головных тем, поэтому чтобы корректно определить штрафную функцию для u-покрытия нужно расширить принадлежность листовых элементов на все узлы дерева. Будем считать, что принадлежности листовых элементов уже нормированы и $\sum\limits_{i\in I} u(i)^2 = 1$. Выберем следующую функцию агрегации для внутренней вершины t: $u(t)=\sqrt{\sum_{i\in I(t)} u(i)^2}$. Итак, ассоциированная с u-покрытием H штрафная функция, которая учитывает важность головных тем, пробелов и выбросов с соответствующими весами 1, λ и γ, будет выглядеть следующим образом:

$$p(H)=\sum_{h\in H\setminus I} u(h)+\lambda\sum_{h\in H\setminus I}\sum_{g\in G(h)} v(g)+\gamma\sum_{h\in H\cap I} u(h), \qquad (1)$$

где λ и γ — гиперпараметры. В статье [3] построен рекурсивный алгоритм, находящий глобальный минимум данной функции. В данной работе мы модифицируем этот метод, исходя из вероятностей приобретения и потерь головных тем в узлах. Данный метод позволяет избавиться от гиперпараметров λ и γ.

Для оценки априорных вероятностей потерь и приобретений для всех узлов, мы применили многократный запуск алгоритма с критерием максимальной экономии (1) примерно 300 раз на случайных наборах по 5000 статей из текстовой коллекции. Для нахождения глобально оптимального решения по критерию максимального правдоподобия используется рекурсивный алгоритм, схожий с рекурсивным алгоритмом для критерия максимальной экономии [3]. Применение нового метода позволяет уточнить и дополнить ранее полученные результаты о тенденциях исследований в области науки о данных.

В работе принимали участие Сузана Насименто (Новый университет Лиссабона, Португалия) и Тревор Феннер (Университет Биркбек, Лондон Великобритания).

Список литературы

[1] *Belkin M.* Laplacian eigenmaps for dimensionality reduction and data representation / M. Belkin, P. Niyogi // Neural Computation. — 2003. — Vol. 15, №6. — P. 1373–1396.

[2] *Chernyak, E.* Refining a taxonomy by using annotated suffix trees and wikipedia resources / E. Chernyak, B. Mirkin // Annals of Data Science. — 2015. — Vol. 2, №1. — P. 61–82.

[3] *Frolov, D.* Parsimonious Generalization of Fuzzy Thematic Sets in Taxonomies Applied to the Analysis of Tendencies of Research in Data Science / D. Frolov, S. Nascimento, T. Fenner, B. Mirkin // Information Sciences. — 2020. — Vol. 512. — P. 595–615.

[4] The 2012 ACM Computing Classification System. — URL: https://www.acm.org/publications/class-2012. — Загл. с титул. экрана.

Библиографическая ссылка

Айрапетян, Ж. С. Метод максимального правдоподобия для обобщения нечетких множеств в таксономиях / Ж. С. Айрапетян, Д. С. Фролов, Б. Г. Миркин // Всероссийская научная конференция «Математические основы информатики и информационно-коммуникационных систем». Сборник трудов. — Тверь : ТвГУ, 2021. — С. 96–101. https://doi.org/10.26456/mfcsics-21-15

Сведения об авторах

1. Айрапетян Жирайр Сережаевич
НИУ «Высшая школа экономики». Исследователь в лаборатории ИССА

Россия, Москва, Покровский бульвар, д.11
E-mail: zhayrapetyan@ithse.ru

2. Фролов Дмитрий Сергеевич
НИУ «Высшая школа экономики». Исследователь МЦАВР

Россия, Москва, Покровский бульвар, д.11
E-mail: dmitsf@gmail.com

3. Борис Григорьевич Миркин
НИУ «Высшая школа экономики». профессор

Россия, Москва, Покровский бульвар, д.11
E-mail: bmirkin@hse.ru