УДК 51-7, 519.1, 519.2

АЛГОРИТМ КЛАСТЕРИЗАЦИИ ГРАФА НА ПРИМЕРЕ ИССЛЕДОВАНИЯ СТРУКТУРЫ ВЗАИМОДЕЙСТВИЯ ПРЕДМЕТНЫХ ОБЛАСТЕЙ В НАУКОМЕТРИЧЕСКОЙ БАЗЕ

Б.Н. Четверушкин

Институт прикладной математики им. М.В. Келдыша РАН

Россия, 125047, Москва, Миусская пл., д.4

E-mail:

М.В. Яшина

Московский автомобильно-дорожный государственный технический университет (МАДИ)

Московский технический университет связи и информатики

Московский авиационный институт (национальный исследовский университет)

Россия, 125319, Москва Ленинградский проспект, 64

Россия, 111024, Москва, Авиамоторная ул, 8а

Россия, Москва, Волоколамское шоссе, 4

E-mail: mv.yashina@madi.ru

А.Г. Таташев

Московский технический университет связи и информатики

Россия, 111024, Москва, Авиамоторная ул, 8а

E-mail: a-tatashev@yandex.ru

Ключевые слова: кластеризация графа, большие данные, непараметрические методы статистического анализа, аттрибутивные коэффициенты корреляции, граф корреляции между научными областями.

Аннотация: В работе рассмотрен алгоритм кластеризации, с помощью которого осуществляется иерархическое разбиение взвешенного графа. На различных уровнях разбиение содержит разное число кластеров (множеств вершин), причем кластеры в разбиении более низкого уровня являются подмножествами кластеров более высоких уровней. Работа алгоритма проанализирована на примере построения формальной классификации областей знаний на основе данных об отнесении индексируемых в наукометрической базе изданий к конкретным областям знаний (предметным областям или их разделам – предметным категориям). В базе каждое издание (журнал, материалы конференции, сборник) отнесены к одной или нескольким категориям. По этим данным вычисляется показатель корреляции между областями знаний, причем ребру графа, вершинам которого соответствуют области знаний, присваивается вес, равный значению этого показателя корреляции. Предлагаемый подход может использоваться для оценки интенсивности взаимодействия между исследователями, работающими в различных областях знаний, и динамики такой интенсивности по годам.

1. Введение

В [1] разрабатывается подход к выявлению групп взаимодействующих между собой исследователей по данным о научных работах, опубликованных этими авторами в соавторстве и индексируемых в базе данных. Этот подход учитывает, что если два исследователя являются соавторами в работе с меньшим числом соавторов, то это является признаком более тесного сотрудничества между этими исследователями, чем при большем числе соавторов. Для оценки взаимодействия между исследователями строится многослойный граф, в фиксированном слое которого вершина соответствует одному из исследователей, а ребра соответствуют научным работам с числом авторов, не превышающим значения, заданного для рассматриваемого уровня. Данный подход обобщает подходы к выявлению групп взаимодействующих авторов на основе только попарных отношений между авторами, каким является, например, подход, изложенный в [2]. В [3] разработан подход к выявлению групп (кластеров) журналов, близких по тематике, на основе числа ссылок в статьях, опубликованных в одном журнале на статью в другом журнала. Результат работы алгоритма показан на примере выявления четырех кластеров в множестве журналов, отнесенных в наукометрической базе «Web of Science» к категории «Transportation (транспорт, перевозки)». Приводимые в [1–3] алгоритмы представляют собой алгоритмы кластеризации графов, т.е. разбиения множества вершин графа (в общем случае взвешенного) на подмножества (кластеры) вершин таким образом, что связи между вершинами внутри кластеров в смысле определенного критерия превышают внешние связи. Известны различные алгоритмы кластеризации и критерии оценки эффективности разбиения, например изложенные в [4–6]. Для обработки больших массивов данных возникает потребность разработки алгоритмов для реализации на высокопроизводительных вычислительных системах [7].

В настоящей статье рассмотрена работа алгоритма кластеризации полносвязаного взвешенного графа, содержащего N вершин. Алгоритм дает разбиения множества вершин на N уровнях. На i-м уровне множество вершин разбито на i кластеров, $i = 1, \ldots, N$. Результат работы алгоритма показан на разбиении множества предметных областей (subject area), выделенных в наукометрической базе Scopus (используются данные портала Scimago [8]). Рассматриваемый алгоритм представляет собой иерархический алгоритм [6] кластеризации, причем в рассматриваемом случае можно считать, что расстояние между парой вершин тем меньше, чем больше вес ребра. Вершины графа соответствуют предметным областям. Веса ребер соответствуют значениям показателя корреляционной связи между предметными областями, вычисляемого по значениям числа индексируемых изданий, отнесенных в базе к обеим областям, числа изданий, не отнесенных к первой области и отнесенных к второй, числа изданий, отнесенных к первой области и не отнесенных к второй, числа изданий, не отнесенных ни к одной из двух областей. В примере использовались данные по журналам, издававшимся в России в 2022 году. Алгоритм может работать на больших массивах данных. Например, можно вершины графа поставить в соответствие предметным категориям (subject category), использовать данные об изданиях различных видов (журналы, материалы конференций, сборники), по различным странам, регионам или по всему миру, исследовать динамику

корреляционных связей между предметными областями (категориями), т.е. динамику интенсивности взаимодействия исследователей, работающих в различных областях знаний, по годам.

2. Описание алгоритма кластеризации и свойства получаемых разбиений

Приведем описание алгоритма кластеризации.

Пусть имеется полносвязный взвешенный граф, содержащий N вершин. Назовем его графом G_1. Работа алгоритма начинается с нижнего – N-го уровня. На первом шаге из графа удаляются все ребра. В результате оказывается, что на N-м уровне каждая вершина представляет собой кластер, состоящий только из этой вершины. При переходе от $(N-i)$-го к $(N-i-1)$-му уровню в граф возвращается ребро с наибольшим весом из ребер, вершины которых находятся в разбиении $(N-i)$-го уровня в различных кластерах, при этом два кластера $(N-i)$-го уровня объединяются в кластер $(N-i-1)$-го уровня, $i = 0, 1, \ldots, N-2$. При равенстве весов двух ребер, предпочтение отдается кластеру, выбираемому по дополнительному признаку. Работа алгоритма заканчивается, если все вершины оказываются содержавшимися в единственном кластере.

Будем называть разбиение (множество кластеров) графа G_1, имеющееся перед переходом от $(N-i)$-го к $(N-i-1)$-му уровню, разбиением $(N-i)$-го уровня или разбиением на $(i+1)$-м шаге, $i = 0, 1, \ldots, N-2$.

Введем граф G_2, наглядно представляющий результат кластеризации графа G_1. Граф G_2 представляет собой дерево высотой $N-1$, корневая вершина которого располагается сверху (на первом уровне) и соответствует кластеру, содержащему все вершины графа G_1. Вершины i-го уровня в дереве G_2 соответствуют кластерам в разбиении i-го уровня для графа G_1, $i = 1, \ldots, N$. Каждая вершина i-го уровня в дереве G_2 имеет одну вершину-потомка среди вершин $(i+1)$-го уровня, соответствующую тому же самому кластеру (множеству вершин) в графе G_1 за исключением вершины, которая соотвтствует кластеру i-го уровня, образованного объединением двух кластеров $(i+1)$-го уровня, и которая соответственно имеет две вершины-потомка среди вершин $(i+1)$-го уровня в графе G_2, $i = 1, \ldots, N-1$.

3. Коэффициенты корреляции качественных признаков

Пусть имеется множества объектов, каждый из которых может обладать обоими качественными признаками A и B, обладать одним из этих признаков и не обладать другим, или не обладать ни одним из этих признаков. Пусть a – число объектов в рассматривамом множестве, обладающих обоими признаками A и B; b – число объектов, не обладающих признаком A и обладающих признаком B; c – число объектов, обладающих признаком A и не обладающих признаком B; d – число объектов, не обладающих ни одним из признаков A и B. Значение, вычисляемое по формуле

$$Phi = \frac{ad - bc}{\sqrt{(a+b)(b+d)(a+c)(c+d)}}, \tag{1}$$

называется коэффициентом контингенции качественных признаков A и B [9]. Другим показателем корреляционной связи качественных признаков является коэффициент ассоциации, вычисляемый по формуле

$$Q = \frac{ad - bc}{(ad + bc)}. \tag{2}$$

4. Иерархический алгоритм кластеризации графа и примнение к анализу больших массивов данных

В базе «Scopus» каждый журнал относится к одной или более чем к одной из 27 предметных областей (subject area) и внутри предметной области к одной к или нескольким предметным категорям. По данным портала «Scimago» в 2022 году в России издавалось 504 журнала, индексировавшихся в базе «Scopus», из которых ни один из этих журналов не был отнесен к предметной категории «Veterinary (ветеринария)». К предметной области «Multidisciplinary (многодисциплинарный)» отнесен один из этих журналов, который отнесен также еще к двум предметным областям. Так как область «Multidisciplinary» не представляет собой конкретную область знаний, эту область не будем рассматривать. Перечислим остальные 25 предметных областей и введем сокращения: 1. Agricultural and Biological Sciences – AaBSc (сельскохозяйственные и биологические науки). 2. Arts and Humanities – AaH (искусствоведение и гуманитарные науки). 3. Biochemistry, Genetics and Molecular Biology – BGaMB (биохимия, генетика и молекулярная микробиология). 4. Business, Management and Accounting – BMaA (бизнес, менеджмент и учет). 5. Chemical Engineering – ChE (химическая инженерия). 6. Chemistry– Ch (химия). 7. Computer Science – CSc (компьютерные науки). 8. Decision Sciences – DSc (науки о принятии решений). 9. Dentistry – D (стоматология). 10. Earth and Planetary Sciences – EaPSc (науки о Земле и п0ланетах). 11. Economics, Econometrics and Finance – EEaF (экономика, экономометрика и финансы). 12. Energy –Ene (энергия). 13. Engineering – Eng (инженерия). 14. Enviromental Science – ESc (наука об окружающей среде). 15. Health Professions – HP (профессии здравохранения). 16. Immunology and Microbiology – IaM (иммунология и микробиология). 17. Materials Science– MSc (наука о материалах). 18. Mathematics – Ma (математика). 19. Medicine – Me (медицина). 20. Neuroscience – Ne (нейронаука). 21. Nursing – Nu (уход за больными). 22. Pharmacology, Toxicology and Pharmaceutics – PhTaPh (фармакология и фармацевтика). 23. Physics and Astronomy – PhaA (физика и астрономия). 24. Psychology – P (психология). 25. Social Sciences – SSc (социальные науки).

Применим алгоритм кластеризации к взвешенному полносвязному графу, построенному по данным о журналах, издававшихся в России в 2022 году. Этот граф содержит 25 вершин, соответствующих перечисленным предметным областям. Вес ребра полагаем равным коэффициенту ассоциации между предметными областями, соответствующими вершинам ребра, который вычисляется по формуле

(2), где a – число журналов, относящихся к обеим областям, b – число журналов, не относящихся к первой области и относящихся ко второй; c – число журналов, относящихся к первой области и не относящихся ко второй; d – число журналов, не относящихся ни к одной из двух областей. При равенстве значений коэффициента ассоциации преимущество предоставляется ребру, для которого больше значение коэффициента контингенции. В пользу предпочтения выбора для рассматриваемых целей коэффициента ассоциации перед коэффициентом контингенции приведем следующий довод. Коэффициент ассоциации равен -1 в том и только в том случае, если ни один журнал не принадлежит одновременно двум рассматриваемым предметным областям ($a = 0$). Коэффициент контингенции при $a = 0$ может иметь различные значения, хотя нет признаков взаимодействия между двумя областями за исключением наличия журналов, не принадлежащих одновременно ни одной из двух областей. Отличие от -1 значения коэффициента ассоциации показывает наличие взаимодействия между областями, заключающееся в наличии журналов, относящихся к обеим областям.

Теорема 1. *Пусть задано значение h меньше веса ребра, добавляемого при переходе от $(25 - i)$-го уровня к $(24 - i)$-му уровню, и больше веса ребра, добавляемого при переходе от $(24 - i)$-го уровня к $(23 - i)$-му уровню, $i = 0, 1, \ldots, 22$. Тогда кластеры разбиения $(25 - i)$-го уровня соответствуют множествам вершин максимальных связных подграфов в графе, получаемом из исходного полносвязного графа удалением ребер с весами, меньшими значения h.*

Следовательно, задание различных значений h задает разбиения множества предметных областей на классы предметных областей на различных уровнях таким образом, что предметные области одного класса взаимодействуют между собой в определенном смысле более интенсивно, чем с предметными областями, содержащимися в других классах.

Теорема 2. *В результате работы алгоритма на рассматриваемом графе получаются разбиения множества предметных областей на 25 уровнях, причем на i-м уровне получается разбиение множества предметных областей на i классов и при переходе от разбиения i-го уровня к разбиению $(i + 1)$-го уровня один из классов предметных областей разбивается на два.*

Теоремы 1 и 2 доказываются с помощью индукции с учетом вида используемого алгоритма.

Проследив шаги образования кластеров в обратном порядке, можно осуществить разбиение классов предметных областей знаний на все более мелкие подклассы.

Полученная в результате работы алгоритма формальная классификация предметных областей может быть представлена в виде дерева (граф G_2 в разделе 2), вершинам которого соответствуют классы предметных областей на соответствующих уровнях.

5. Заключение

Рассмотрен алгоритм кластеризации графа, дающий разбиения множества вершин графа на различных уровнях. Работа алгоритма проанализирована на примере построения условной классификации областей знаний с использованием данных о том, к каким областям знаний относятся в наукометрической базе

индексируемые издания (каждое издание отнесено в базе к одной или нескольких областей знаний).

Список литературы

1. Vasilyeva E., Kozlov A., Alfaro-Bittner K., Musatov D., Raigorodskii A.M., Perc M., Boccaletti S. Multilayer representation of collaboration networks with higher-order interactions // Scientific Reports. 2021. Vol. 11, No. 1. P. 5666.
2. Newman M.E. The structure of scientific collaboration networks // Proceedings of the national academy of sciences. 2001. Vol. 98, No. 2. P. 404–409.
3. Haghani M., Bliemer M.C.J. Structure and temporal evolution of transportation literature. 2021. arXiv 2107.12639. 2021
4. Fortunato S. Community detection in graphs // Physics Reports. 2010. Vol. 486. P. 75–144.
5. Miasnikof P., Prochorenkov L., Shestopalov A.Y., Raigorodski A. A statistical test of heterogenous subgraph densities to assess clusterability // Proceedings of the 13th International Conference LION 13 Learning and Intelligible Optimization. Chania, Crete, Greece, May 27–31. Cham: Springer, 2019. P. 17–29.
6. Воронцов К.В., Лекции по алгоритмам кластеризации многомерного шкалирования. Курс лекций. 2007. hhttps://habr.com/ru/articles/101338
7. Четверушкин Б.Н., Савельев А.В., Савельев В.И. Моделирование задач магнитной гидродинамики на высокопроизводительных вычислительных системах // Математическое моделирование. 2020. Т. 32, № 12. P. 3-–13.
8. Scimago Journal & and Country Rank. `https://www.scimagojr.com` (дата обращения: 29.12.2023).
9. Елисеева И.И., Юзбашев М.М. Общая теория статистики. М.: Финансы и статистика, 2006. 656 с.