АКТУАЛЬНЫЕ ПРОБЛЕМЫ ИНФОРМАТИКИ

УДК 004.048

Применение методов машинного обучения для анализа научных публикаций и оценка их эффективности

Малашков Валентин Борисович[a], Шульдова Светлана Георгиевна[b], Лапицкая Наталья Владимировна[c]

[a] ООО «Фло Хелпс Бел», инженер-программист, магистр, malashkovv@gmail.com
[b] Белорусский университет информатики и радиоэлектроники, кандидат технических наук, доцент, доцент кафедры программного обеспечения информационных технологий, shsg@bsuir.by
[c] Белорусский университет информатики и радиоэлектроники, кандидат технических наук, доцент, заведующий кафедрой программного обеспечения информационных технологий, lapan@bsuir.by

Аннотация
В статье рассмотрены методы машинного обучения, используемые для анализа научных публикаций с целью определения перспективных научных направлений или детектирования сообществ ученых. Осуществлена их программная реализация, а также выполнена оценка эффективности с точки зрения качества кластеризации.

Ключевые слова: публикация, цитирование, граф, кластеризация, векторное представление, метод, анализ.

Веб: http://library.miu.by/journals!/item.science-xxi/issue.10/article.1.html

Поступила в редакцию: 07.12.2021

Application of machine learning methods for analysis of scientific publications and rating of their efficiency

Malashkou Valiantsin[a], Shuldava Sviatlana[b], Lapitskaya Nataliya[c]

[a] OOO «Flo Helps Bel», Software engineer, Master's degree holder, malashkovv@gmail.com
[b] Belarusian State University of Informatics and Radioelectronics, PhD in Technical Sciences, Associate Professor, Associate Professor in the Department of Software of Information Technologies, shsg@bsuir.by
[c] Belarusian State University of Informatics and Radioelectronics, PhD in Technical Sciences, Associate Professor, Head of the Department of Software of Information Technologies, lapan@bsuir.by

Abstract
The article discusses machine learning methods that are used to analyze scientific publications for the definition of hopeful scientific research or the detection of communities of scientists. Their software implementation has been carried out, and their efficiency has been assessed in terms of the quality of clustering.

Keywords: publication, citation, graph, clustering, presentation, method, analysis.

Web: http://library.miu.by/journals!/item.science-xxi/issue.10/article.1.html

Received: 07.12.2021

Введение

В настоящее время отдельному ученому или научному работнику очень сложно охватить всевозрастающие потоки данных научных исследований в своей области, а если учитывать междисциплинарное направление, то объем исходных данных многократно увеличивается. В этой связи для эффективной обработки и классификации научных публикаций, выявления актуальности, перспективности различных научных направлений, детектирования научных сообществ, а также предотвращения массового тиражирования работ, лишенных ценности, широко используются методы интеллектуального анализа, исследованию и применению которых посвящена настоящая работа.

1. Методы и модели анализа

Определение актуальных направлений научных исследований и детектирование научных сообществ предполагают разбиение множества документов из одной предметной области (кластеризация) на близкие по смыслу подмножества (кластеры). Задача кластеризации может быть решена с помощью методов анализа текстовой информации, в качестве которой есть возможность использовать как полные тексты статей, так и их аннотации или даже названия. Некоторые из этих методов основаны на взаимной встречаемости слов в текстах статей [1]; другие строят тематическую модель текстов статей и на ее основе производят кластеризацию [2].

Кластеризация научных статей может быть выполнена на основе графа цитирования $G = (V, E)$, в котором вершины $V = \{v_1, v_2, \ldots, v_n\}$ соответствуют статьям, а ребра – отношениям цитирования $E \subseteq V \times V$, $e = (v_i, v_j) \in E$, $i \neq j$, так как самоцитирование исключено [2]. В этом случае статьи образуют те же кластеры, что и соответствующие им вершины графа цитирования.

Один из возможных подходов решения задачи кластеризации заключается в нахождении векторного представления статьи (сопоставлении каждой статье вещественного вектора фиксированной длины) и кластеризации этих векторов. В этом случае в один кластер попадают те статьи, чьи векторы также попали в один кластер.

В качестве метода построения векторного представления текста используется метод Word2vec, разработанный группой исследователей Google для анализа семантики естественных языков [3]. Метод Word2vec строит векторное представление слова на основе большого набора текстов. В результате каждое слово представляется в виде вектора, состоящего из координат слов. Два слова имеют близкие векторы в том случае, если они встречаются в текстах рядом с одними и теми же словами.

На основе Word2vec разработаны методы, вычисляющие как векторное представление текстов, так и векторное представление вершин графа цитирования. Node2vec – метод, который преобразует вершины графа (или сети) в векторные представления [4]. Это представление в некотором смысле сохраняет структуру исходной сети, так что вершины, тесно связанные, имеют векторные представления, угол между которыми меньше, чем у векторных представлений слабосвязанных вершин.

Использование методов построения векторного представления вершин графа, основанных на Word2vec, в задаче кластеризации является более эффективным, чем использование классического алгоритма кластеризации графа.

Авторами для анализа научных направлений и сообществ на основании текстов публикаций и ассоциированной с ними информации рассмотрены два подхода.

Первый подход основан на применении лувенского метода [5], который относится к агломеративной иерархической кластеризации и основан на максимизации модулярности, применяемой для оценки качества разбиения графа. В качестве способа задания графа цитирования $G = (V, E)$ используется матрица смежности $A(G)$. Под элементом матрицы A_{ij} понимается вес ребра, соединяющего вершины i и j. В качестве функции для подсчета весов ребер в графе цитирования используется мера схожести текстов научных публикаций.

Модулярность – это мера, которая равна доле ребер от общего числа ребер, которые попадают в данные группы, минус ожидаемая доля ребер, которые попали бы в те же группы, если бы они были распределены случайно. Данная мера считается по следующей формуле:

$$Q = \frac{1}{4|E|}\sum\left(A_{ij} - \frac{k_i k_j}{2|E|}\right)\delta_{c_i c_j}, \qquad (1)$$

где $|E|$ – количество ребер графа;

A_{ij} – элемент матрицы смежности;

k – степень узла;

$\delta_{c_i c_j}$ – дельта-функция, которая обозначает принадлежность узлов i и j к одному сообществу.

Оптимизация выполняется в два этапа. Метод ищет небольшие сообщества, оптимизируя модулярность локально (используя метод округления), а затем создает иерархическое подсообщество, дочерними элементами которого являются два сообщества, обнаруженные на первом шаге. Этот процесс повторяется до тех пор, пока не будет достигнут максимум модулярности.

Задача поиска выделения сообществ в графе сводится к поиску таких c_i, которые максимизируют значение модулярности.

Косинусное расстояние – это мера сходства между двумя векторами внутреннего пространства произведения, которая измеряет косинус угла между ними. Определяется как отношение скалярного произведения векторов к произведению их длин [6]:

$$d(\boldsymbol{p}, \boldsymbol{q}) = 1 - \frac{\boldsymbol{p} * \boldsymbol{q}}{|p||q|}, \quad (2)$$

где $\boldsymbol{p}$ и $\boldsymbol{q}$ – вектора.

Если обратить внимание на формулу (2), то можно отметить, что от единицы отнимается значение косинуса угла между векторами. Такой подход называется «расстоянием», а значит, семантически схожие векторы будут давать значение, близкое к 0.

Последовательность анализа следующая:

1. Получение векторного представления на основе текстов научных публикаций.

2. Построение графа цитирования.

3. Подсчет весов ребер графа цитирования на основе меры схожести векторного представления текстов научных публикаций.

4. Применение алгоритма для обнаружения сообществ и выявления научных направлений.

Второй подход заключается в получении векторного представления вершин графа цитирования и применении алгоритма кластеризации. Порядок выполнения представлен ниже.

1. Получение векторного представления на основе текстов научных публикаций.

2. Построение графа цитирования.

3. Получение векторного представления для вершин графа цитирования.

4. Применение алгоритма кластеризации поверх векторного представления вершин графа цитирования и текстов научных публикаций.

Алгоритм *k*-средних – это итерационный алгоритм, который пытается разбить набор данных на определенные *k*-подгруппы (кластеры), где каждая точка данных принадлежит только одной группе. Алгоритм пытается сделать точки данных внутри кластера как можно более похожими, сохраняя при этом кластеры как можно более разными (далекими). Он назначает точки данных кластеру таким образом, чтобы сумма квадратов расстояния между точками данных и «центроидом» кластера (среднее арифметическое всех точек данных, принадлежащих этому кластеру) была минимальной. Чем меньше вариаций имеется внутри кластеров, тем более однородными (похожими) являются точки данных внутри одного и того же кластера [7].

2. Источники данных для анализа

Источники данных анализа можно разделить на две категории:

– поисковые системы и архивы научных публикаций;

– открытые наборы данных.

Наиболее известными научными поисковыми системами, которые объединяют электронные базы данных научного цитирования, являются Web of Knowledge, Scopus, Google Scholar и Российский индекс цитирования (РИНЦ).

Web of Knowledge (WoK) – это интегрированная web-платформа, созданная компанией Thomson Reuters с целью фильтрования информационного потока научных публикаций и событий (конференций, патентов, сайтов и данных, доступных в открытом доступе). В зависимости от оплаченной научной, учебной или экспертной организацией подписки на продукцию WoK спектр доступных для работы баз данных может быть различным.

База данных Scopus позиционируется издательской корпорацией Elsevier как крупнейшая в мире универсальная реферативная база данных с возможностями отслеживания научной цитируемости публикаций.

Google Scholar (Google Академия) – это бесплатная поисковая система, которая индексирует полный текст научных публикаций всех форматов и дисциплин в большинстве рецензируемых журналов Европы и Америки, доступных в физической или онлайн-копиях.

Российский индекс научного цитирования – библиографическая база данных научных публикаций российских ученых. Размещается на сайте научной электронной библиотеки elibrary.ru. РИНЦ помогает составить картину публикационной деятельности на основе внутрироссийского мониторинга.

Помимо систем, где необходимо самостоятельно извлекать данные, существуют готовые наборы данных. Они обладают разным набором атрибутов и собраны из разных источников под различные исследовательские задачи. Среди готовых наборов данных можно выделить следующие:

– PubMed;

– CORE;

– Semantic Scholar;

– DBLP.

PubMed первоначально являлся поисковой системой, ориентированной на статьи медицинской тематики. В настоящее время она насчитывает более 32 миллионов статей, о которых содержаться следующие данные:

– авторы;

– краткое содержание;

– цифровой идентификатор объекта (DOI);

– название;

– цитирующие статьи.

Набор данных CORE доступен в двух экземплярах. Первая база данных содержит 123 миллиона строк метаданных о статьях, 85,6 миллиона из них – вместе с кратким содержанием. Архив занимает 127 Гб дискового пространства. Вторая версия дополнительно содержит 9,8 миллиона статей с полным текстом, что в свою очередь увеличивает объем занимаемого дискового пространства до 330 Гб.

Сайт Semantic Scholar предоставляет доступ как посредством API с определенным количеством запросов в минуту, так и в виде архива. В данных присутствуют:

– информация об авторах;

– цитирования;

– источник данных о статье;

– краткое содержание;

– DOI;

– уникальный идентификатор в системе Semantic Scholar;

– уникальные идентификаторы из других систем, если таковые имеются (например, Microsoft Academic Search).

Поисковая система DBLP специализируется на статьях, связанных с компьютерными науками. DBLP предоставляет открытую библиографическую информацию по основным журналам и трудам в области компьютерных наук. Первоначально этот набор данных был создан в университете Трира в 1993 году.

Самостоятельно сайт не предоставляет готовое хранилище данных, но на его основе были собраны наборы данных сообществ исследователей, например, набор данных от команды Aminer [4]. Этот набор данных имеет различные версии и, соответственно, разное количество статей, а также атрибутов, ассоциированных с ними.

Таким образом, перспективным источником данных является набор данных DBLP, так как он имеет гибкую вариативность в размерах и количестве соединений в графе цитирования, базовый набор наукометрических показателей, краткое содержание, название и название сборника публикаций.

3. Инструменты реализации

Исходным набором данных является DBLP от AMiner [8], который содержит 3 079 007 статей, а также 25 166 994 соединений посредством ссылок на источники. В таблице 1 представлены атрибуты, которые доступны для анализа.

Таблица 1 – Формат данных DBLP

Атрибут	Описание
abstract	Краткое содержание
authors	Список авторов
n_citations	Количество цитирований на данную статью
references	Ссылки на источники, которые представлены в виде списка «id» других строк данного датасет
title	Название
venue	Сборник публикации или конференция
year	Год публикации
id	Уникальный идентификатор UUID

Код реализации экспериментальной части написан на языке программирования Python версии 3.7. Реализация алгоритма машинного обучения Word2vec присутствует в стороннем пакете Genism, который является одним из самых популярных средств для работы с естественным языком как в промышленной разработке, так и в исследовательской деятельности. В реализации алгоритма Word2vec от Gensim присутствуют все необходимые гиперпараметры для точечной настройки работы алгоритма.

Полученные векторы представления текстов использованы в связке с лувенским методом, алгоритмом машинного обучения Node2vec и алгоритмом кластеризации *k*-средних.

В качестве рассматриваемых программных реализаций лувенского метода были отобраны несколько сторонних, доступных на языке программирования Python:

– библиотека Cugraph;

– библиотека Cdlib.

Для представления графа цитирования используется сторонняя библиотека Networkx, так как является стандартом в области обработки графа на языке программирования Python и все остальные инструменты чаще всего уже имеют интеграцию с данным программным средством.

Далее с помощью сторонней библиотеки Sklearn для языка программирования Python происходит подсчет весов ребер на основе «косинусного расстояния». Также был опробован вариант, где оценка схожести текстов вычисляется как экспонента от функции косинусного расстояния. Данный вариант был использован с целью сгладить частотное распределение весов внутри графа, уровень сглаживания контролируется с помощью параметра σ. Функция вычисляется следующим образом:

$$\frac{e^s}{\sigma}, \tag{3}$$

где s – оценка схожести на основании косинусного расстояния;

σ – параметр сглаживания.

Для реализации алгоритма Node2vec на языке программирования Python использована одноименная библиотека. Кластерный анализ выполняется с помощью алгоритма *k*-средних в реализации библиотеки Sklearn, а основным входным параметром является количество кластеров.

4. Результаты исследования

Для оценки результатов работы всех вариантов алгоритмов используется граф размером в 20 000 вершин.

При оптимизации алгоритма лувенского метода использовался метод оценки модулярности сети при разбиении на отдельные кластеры по формуле (1). По опыту исследователей ориентироваться на данную метрику стоит с осторожностью, так как она не совсем четко отражает конечный результат, но в то же время полезна для понимания того, насколько хорошо сеть графа цитирования разбивается на отдельные подграфы [9]. Чем ближе значение к единице, тем лучше, так как лувенский метод оптимизирует именно эту целевую функцию.

Результаты оценки модулярности для лувенского метода без весов, с количеством кластеров и значением параметра разрешения представлены в таблице 2.

Таблица 2 – Оценка модулярности лувенского метода

Разрешение	Модулярность	Количество кластеров
1,0	0,772	44
0,1	0,683	222

Как видно из таблицы 2, с уменьшением параметра «Разрешение», увеличивается количество кластеров, но при этом падает значение метрики «Модулярность».

Результаты оценки модулярности для лувенского метода в сочетании с векторным представлением текста краткого содержания, полученные в результате модели Word2vec, представлены в таблице 3.

С помощью оценки модулярности сложно сделать вывод, насколько хорошо работает тот или иной метод, но при этом можно увидеть, что в некоторых случаях добавление веса в ребра графа на основе оценки схожести текстов кратких содержаний научных публикаций улучшает метрику либо метрика остается примерно на том же значении, но увеличивается количество кластеров.

Таблица 3 – Оценка модулярности лувенского метода в сочетании с мерами схожести кратких содержаний научных публикаций на основе модели Word2vec

Мера схожести текста	Разрешение	Модулярность	Количество кластеров
Косинусное расстояние	1,0	0,769	38
Косинусное расстояние	0,1	0,682	217
Формула (3) при $\sigma = 1$	1,0	0,769	36
Формула (3) при $\sigma = 1$	0,1	0,680	218
Формула (3) при $\sigma = 0{,}1$	1,0	0,764	55
Формула (3) при $\sigma = 0{,}1$	0,1	0,662	250

Для оценки кластеризации всех ранее описанных подходов будут использованы две основные метрики [8]:

– ARI (Adjusted Rand Index);

– AMI (Adjusted Mutual Info).

ARI является мерой расстояния между различными разбиениями выборки и принимает значения в диапазоне [–1, 1]. Отрицательные значения соответствуют «независимым» разбиениям на кластеры, значения, близкие к нулю,– случайным разбиениям, положительные же значения говорят о том, что два разбиения схожи (совпадают при ARI = 1). AMI определяется с использованием функции энтропии, интерпретируя разбиения выборки как дискретные распределения (вероятность отнесения к кластеру равна доле объектов в нем). Он принимает значения в диапазоне [0, 1]. Значения, близкие к нулю, говорят о независимости разбиений, а близкие к единице – об их схожести (совпадении при AMI = 1). Обе метрики реализованы в библиотеке Sklearn для языка программирования Python.

Сочетание векторов, полученных с помощью алгоритмов Word2vec и Node2vec, как входных данных для алгоритма кластеризации показало лучший результат с точки зрения метрик ARI и AMI.

На рисунке представлен пример кластеризации выигрышным подходом для подграфа на 100 вершин.

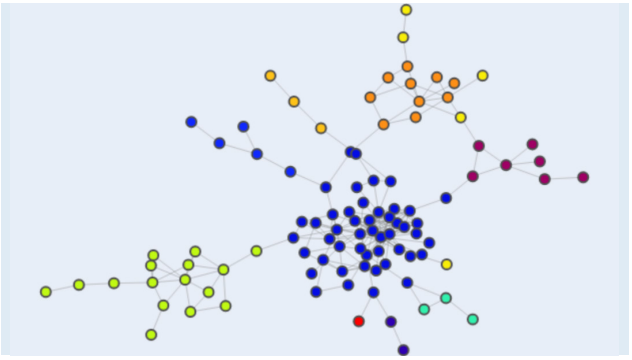

Рисунок – Пример разделения сети научных публикаций с помощью сочетания алгоритмов Node2vec и Word2vec для подграфа на 100 вершин

Полученные кластеры характеризуют области, которыми занимается значительное число исследователей.

Заключение

На основании оценки эффективности двух разных подходов можно сделать вывод, что оба подхода позволяют сформировать кластеры документов для последующей обработки и извлечения ключевой информации для определения актуальных научных направлений. Наилучший результат показал метод кластеризации на основе векторного представления вершин и текста; подход на основе лувенского метода, где веса являлись оценкой схожести текстов научных публикаций, показал сравнимый результат.

Можно сделать вывод и о том, что огромную часть контекстной информации о научной публикации несет в себе информация о ссылках на источники.

Результаты исследования и используемые инструменты полезны магистрантам и аспирантам при выборе темы исследования и обзоре источников.

ЛИТЕРАТУРА / REFERENCES

1. Review of automatic text summarization techniques & methods / A. P. Widyassari [et al.] // Journal of King Saud University – Computer and Information Sciences. – 2020.

2. Marchionini, G. Exploratory search: from finding to understanding / G. Marchionini // Communications ofthe ACM.– 2006. – Vol. 49, № 4. – P. 41–46.

3. Янина, А.О. Мультимодальные тематические модели для разведочного поиска в коллективном блоге / А.О. Янина, К.В. Воронцов // Машинное обучение и анализ данных. – 2016. – Т. 2, № 2. – С. 173–186.

Yanina, A.O. Mul'timodal'nyye tematicheskiye modeli dlya razvedochnogo poiska v kollektivnom bloge / A.O. Yanina, K.V. Vorontsov // Mashinnoye obucheniye i analiz dannykh. – 2016. – T. 2, № 2. – P. 173–186.

4. Community detection in complex networks using Node2vec with spectral clustering / F. Hu [et al.] // Physica A: Statistical Mechanics and its Applications. – 2020. – № 545. – P. 623–633.

5. Fast unfolding of community hierarchies in large networks [Electronic resource] / V.D. Blondel [et al.] // ArXiv. – Mode of access: https://arxiv.org/pdf/0803.0476.pdf. – Date of access: 16.05.2021.

6. KMeans [Electronic resource]. – Mode of access: https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html. – Date of access: 16.05.2021.

7. ArnetMiner: Extraction and Mining of Academic Social Networks / J. Tang [et al.] // Proceedings of the Fourteenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (SIGKDD'2008).– 2008.– P. 990–998.

8. A Review of Clustering Techniques and Development / A. Saxena [et al.] // Neurocomputing. – 2017. – № 267.

9. Clustering Metrics [Electronic resource]. – Mode of access: https://scikit-learn.org/stable/modules/classes.html#clustering-metrics. – Date of access: 16.05.2021.