УДК 001

БЛИЗОСТЬ ПРОФИЛЕЙ НАУЧНЫХ ПУБЛИКАЦИЙ В СЕТИ СОАВТОРСТВА НА ПРИМЕРЕ ИПУ РАН

Д.А. Губанов
Институт проблем управления им. В.А. Трапезникова РАН
Россия, 117997, Москва, Профсоюзная ул., 65
E-mail: dmitry.a.g@gmail.com

В.С. Мельничук
Институт проблем управления им. В.А. Трапезникова РАН
Россия, 117997, Москва, Профсоюзная ул., 65
E-mail: vs.melnichuk09@gmail.com

Ключевые слова: сеть публикаций, профиль публикации, теория управления.

Аннотация: Исследуется взаимосвязь между профилями научных публикаций, связанных отношением соавторства, в области теории управления. Описывается базовый алгоритм расчета профиля научной публикации. Сформулированы и исследованы вопросы относительно близости публикаций в сети соавторства, на основе полученных результатов предложены подходы к прогнозированию профилей публикаций. Описанные подходы апробируются в Информационной системе автоматизации научной деятельности (ИСАНД), которая разрабатывается в ИПУ РАН.

1. Введение

В настоящее время актуальными проблемами, требующими решения, являются автоматизированный анализ научных публикаций и поддержка принятия научно-организационных решений в научных учреждениях, редакциях журналов и других научных организациях. Эти задачи включают в себя направление материалов на рецензирование [4], оценку новизны статьи [3], а также поддержку деятельности ученых, такой как подбор статей по определенной тематике [2, 5] и др. [6, 7]. Для эффективного решения этих задач необходимо определение тематики научного текста. Это, в свою очередь, требует позиционирования научных объектов в тематическом пространстве науки, которое мы называем профилем научного объекта. Ранее в рамках ИСАНД был разработан специализированный классификатор в области теории управления. Этот классификатор был использован нами для построения профилей научных объектов на основе текстов публикаций [1]. В то же время в работах [8, 9, 10] показано, что учет сетевых связей между объектами позволяет добиться большей точности классификации объектов в различных предметных областях. Поэтому можно ожидать повышения качества профилей научных публикаций, принимая в расчет не только тексты публикаций, но и сетевую структуру связей между ними.

В разделе 2 кратко описан базовый подход к оценке профилей публикаций, основанный на анализе текстов публикаций. В разделе 3 изложены результаты анализа использования информации о связях между публикациями для оценки профилей публикаций. В Заключении приводится их краткое обсуждение и перспективы развития.

2. Исходные данные

2.1. Расчет профилей публикаций

Кратко приведем, следуя [1], основные обозначения и базовый алгоритм расчета профилей научных публикаций. В работе [1] предложен классификатор ИСАНД

(онтология теории управления), существенным отличием которого от известных классификаторов является его многомерность. Этот специализированный классификатор является системой координат тематического пространства. Характеристикой объекта в пространстве является вектор, называемый *профилем*. Онтология системы ИСАНД имеет четырехуровневую структуру, уровни которой (кроме нижнего) представляют собой дерево. Уровни нумеруются числами от 0 до 4. Уровень 0 содержит 3 фиксированные вершины «Математический аппарат», «Проблемная область» и «Сфера применения», отражающие различные аспекты научных исследований. Каждая из вершин уровня 0 является корнем тематического поддерева, раскрывающего ее содержание.

Множество вершин 1-го уровня, называемых темами, обозначается через $V = \{v_1, \ldots, v_n\}$. При этом i-я вершина 1-го уровня связана с множеством $V_i = \{v_{i1}, \ldots, v_{in_i}\}$ вершин 2-го уровня – подтем. Общее число подтем $m = \sum_{i \in N} n_i$. Третий уровень – это вершины-термины, характеризующие подтемы. Обозначим L – множество публикаций, Δ_{lij} – сумма числа вхождений в l-ю публикацию базовых терминов из ij-й подтемы. Каждая публикация может характеризоваться различными стохастическими векторами. Стохастический профиль второго уровня публикации l:

$$x_l = (x_{l1}, \ldots, x_{lij}, \ldots, x_{lnm}),$$

где $x_{lij} = \frac{\Delta_{lij}}{\sum_{i \in N} \sum_{j \in N_i} \Delta_{lij}}, l \in L, j \in N_i, i \in N.$

Профиль первого уровня публикации l:

$$X_l = (X_{l1}, \ldots, X_{li}, \ldots, X_{ln}), \text{ где } X_{li} = \sum_{j \in N_i} x_{lij}, \ l \in L, \ i \in N.$$

2.2. Построение сети публикаций

В данной работе мы рассматриваем сеть соавторства публикаций. В ней узлами являются публикации, а ненаправленная дуга между двумя узлами означает наличие совместных авторов. В дальнейшем будем рассматривать сеть публикаций, авторами которых являются сотрудники ИПУ РАН. Такая сеть состоит из 14 тыс. узлов и 866 тыс. связей, число компонент связности – 29, наибольшая компонента содержит 94 % от всех публикаций. Средняя степень узла – 123, медиана – 98.

3. Результаты

Для каждой публикации в сети соавторства рассчитан профиль с использованием базового метода (см. раздел 2.1). Несмотря на неплохое в целом качество работы такого метода, следует учитывать и его ограничения, в частности: неизбежные пробелы в тематическом классификаторе (например, те или иные термины могут быть посчитаны слишком общими для включения в классификатор, или слишком специфичными, неточными), а также неоднородность качества составления аннотаций публикаций. Все это может привести к формированию недостаточно точных или даже ошибочных профилей публикаций. В то же время учет дополнительных метаданных (информации о соавторстве) может привести к улучшению качества построения профилей. Мы последовательно рассмотрели ряд вопросов, ответы на которые могут улучшить качество:

1) В какой мере профиль заданной публикации отличается от профилей всех прочих публикаций в выборке?
2) Обладает ли профиль данной публикации сходством с профилями других работ, опубликованных ее авторами в близкий временной промежуток?
3) Определяется ли степень сходства профилей публикаций сходством коллектива авторов?
4) Следует ли из близости содержимого текстов близость профилей соответствующих публикаций?

Для оценки сходства профилей используется расстояние, основанное на манхэттенской метрике (расстояние лежит в промежутке от 0 до 1, см. подробнее [1]).

Рассматриваются публикации, которые содержат не менее *r* терминов из тематического классификатора.

Анализ показывает, что расстояние для случайно выбранных публикаций является практически максимальным (см. распределение на рис. 1 и таблицу 1, в которой приведены среднее расстояние и медиана), таким образом набор публикаций не является однородным и охватывает разные темы теории управления (Вопрос 1).

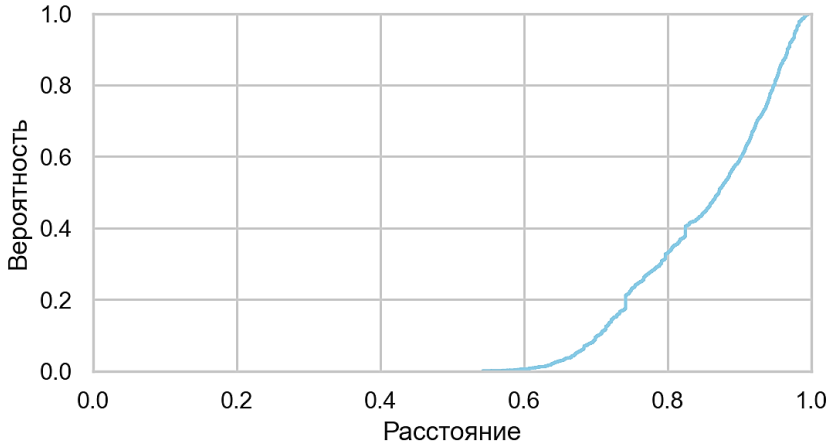

Рис. 1. Эмпирическая функция распределения расстояний между профилями случайно выбранных публикаций

Из таблицы 1 следует, что: 1) если известно, то публикации связаны отношением соавторства, то расстояние между ними существенно меньше, чем в случае случайной пары публикаций (среднее 0,60, медиана 0,59) (Вопрос 2); 2) чем меньше между этими публикациями временной промежуток, тем больше между ними сходства (Вопрос 2).

Таблица 1. Расстояния между парами публикаций (среднее и медиана).

Число терминов	Случайные узлы	Смежные узлы	Смежные узлы (3 года)	Смежные узлы (1 год)
$r = 3$	0,91 (0,95)	0,68 (0,75)	0,67 (0,71)	0,66 (0,69)
$r = 5$	0,90 (1,00)	0,60 (0,59)	0,59 (0,57)	0,58 (0,56)

Для прогноза профиля заданной публикации рассмотрим следующие способы агрегирования соседних профилей:

1) среднее соседних профилей;
2) взвешенная сумма соседних профилей (вес – сходство коллективов авторов);
3) взвешенная сумма соседних профилей (вес – сходство коллективов авторов) с отсечением непохожих по содержанию соседних профилей.

Введем обозначения:

- $K(l)$ – множество авторов l-ой публикации, $l \in L$,
- $t(l)$ – год публикации l-ой публикации,
- $T = [t_1; t_2] = [t(l) - \delta; t(l) + \delta]$ – промежуток времени,
- $|K(m) \cap K(l)|/|K(m) \cup K(l)|$ – мера сходства публикаций l и m по составу авторов,
- x_l – базовый стохастический профиль публикации l,
- $\hat{x}_l$ – прогноз стохастического профиля публикации l.

Тогда второй способ прогноза определяется так:

$$\hat{x}_l = \sum_{m \in L} \frac{|K(m) \cap K(l)|}{|K(m)|} x_m \Big/ \sum_{m \in L} \frac{|K(m) \cap K(l)|}{|K(m)|}.$$

С учетом полученных результатов примем $\delta = 1$. Для оценки сходства содержимого текстов будем применять языковые нейросетевые модели sciBERT (для русскоязычных текстов, см. https://github.com/allenai/scibert) и ruSciBERT (для англоязычных текстов, см. https://huggingface.co/ai-forever/ruSciBERT). Модели используются для преобразования текстов публикаций в векторное представление, для оценки сходства

применяется косинусное расстояние. Результаты анализа представлены в таблице 2. Наилучшие результаты показывает 3 способ прогноза профиля публикации. Кроме того, чем ближе содержимое текстов публикаций, тем ближе профили публикаций (Вопрос 3); чем больше степень совпадения коллективов авторов, тем ближе профили публикаций (Вопрос 4).

Таблица 2. Оценки качества способов прогноза профилей.

Число терминов	Окружение	Вес окружения	Порог BERT 0,5	Порог BERT 0,1
$r = 3$	0,59 (0,59)	0,56 (0,55)	0,55 (0,54)	0,48 (0,45)
$r = 5$	0,53 (0,50)	0,50 (0,46)	0,49 (0,45)	0,43 (0,39)

4. Заключение

Проведен анализ сети публикаций, связанных отношением соавторства: рассмотрены вопросы взаимосвязи профилей публикаций сети и возможности прогнозирования профилей публикаций на основе связей соавторства и текстов публикаций. Оказалось, что профили публикаций, соединенных связью соавторства, обладают значительно большей степенью сходства, чем профили случайно выбранных публикаций. Кроме того, на близость профилей также влияет временной промежуток между публикациями (чем меньше, тем ближе), соотношение коллектива авторов (чем ближе к 1, тем ближе) и различия в содержании текстов публикаций (чем меньше, тем ближе). На основе этих результатов предложены методы прогнозирования профилей, показавшие свою адекватность.

Перспективным направлениями дальнейших исследований являются разработка методов построения профилей публикаций с учетом полученных знаний о сетевой структуре, а также сравнение методов с SOTA-решениями в данной области на примере общедоступных графовых датасетов.

Список литературы

1. Губанов Д.А., Кузнецов О.П., Суховеров В.С., Чхартишвили А.Г. О построении профилей в тематическом пространстве теории управления // Материалы 9-й Международной конференции «Знания-Онтологии-Теории» (ЗОНТ-2023, Новосибирск). Новосибирск: Институт математики им. С.Л. Соболева СО РАН, 2023. С. 89–94.
2. Крыжановская С.Ю., Власов А.В., Еремеев М.А., Воронцов К.В. Полуавтоматическая суммаризация тематических подборок научных публикаций: задачи и подходы // Тезисы докладов 20-й Всероссийской конференции с международным участием «Математические методы распознавания образов». М.: Российская академия наук, 2021. С. 333-338.
3. Shibayama S., Yin D., Matsumoto K. Measuring novelty in science with word embedding. PLoS ONE 2021. Vol. 16, No. 7. P. e0254034. https://doi.org/10.1371/journal.pone.0254034.
4. Yuan W., Liu P., Neubig G. Can we automate scientific reviewing? // Journal of Artificial Intelligence Research. 2022. Vol. 75. P. 171-212.
5. Cachola I., Lo K., Cohan A., Weld D. TLDR: Extreme Summarization of Scientific Documents. // Findings of the Association for Computational Linguistics: EMNLP. 2020. P. 4766-4777.
6. Peng Bao, Weihui Hong, Xuanya Li. Predicting Paper Acceptance via Interpretable Decision Sets // Companion Proceedings of the Web Conference 2021 (WWW '21). Association for Computing Machinery, New York, NY, USA. P. 461-467. https://doi.org/10.1145/3442442.3451370.
7. Kasanishi T., et al. SciReviewGen: A Large-scale Dataset for Automatic Literature Review Generation. arXiv preprint arXiv:2305.15186. 2023.
8. Hasegawa T., Arvidsson H., Tudzarovski N., Meinke K., Sugars R.V., Ashok Nair A. Edge-Based Graph Neural Networks for Cell-Graph Modeling and Prediction // Frangi A., de Bruijne M., Wassermann D., Navab N. (Eds.) Information Processing in Medical Imaging. IPMI 2023. Lecture Notes in Computer Science. Cham: Springer. Vol. 13939. https://doi.org/10.1007/978-3-031-34048-2_21.
9. Xiong C., Li W., Liu Y., Wang M. Multi-Dimensional Edge Features Graph Neural Network on Few-Shot Image Classification // IEEE Signal Processing Letters. 2021. Vol. 28. P. 573-577. doi: 10.1109/LSP.2021.3061978.
10. Faber L., Lu Y., Wattenhofer R. Should Graph Neural Networks Use Features, Edges, Or Both? https://arxiv.org/abs/2103.06857. 2021.
11. Zhou J., et al. Graph neural networks: A review of methods and applications // AI Open. 2020. P. 57-81.