УДК 519.1

ПОСТРОЕНИЕ ПРОФИЛЕЙ НАУЧНЫХ ПУБЛИКАЦИЙ НА ОСНОВЕ ТЕКСТОВ И СВЯЗЕЙ СОАВТОРСТВА (НА ПРИМЕРЕ ТЕОРИИ УПРАВЛЕНИЯ И ЕЕ ПРИЛОЖЕНИЙ)

Д. А. Губанов*, В. С. Мельничук**

*,**Институт проблем управления им. В.А. Трапезникова РАН, г. Москва
**МГТУ им. Н.Э. Баумана, г. Москва

*✉ dmitry.a.g@gmail.com, **✉ vs.melnichuk09@gmail.com

Аннотация. Расчет профилей научных публикаций играет ключевую роль в систематизации научных знаний и поддержке принятия научных решений. Предложен метод формирования профилей публикаций в области теории управления, основанный на интеграции анализа текстов и анализа сетей соавторства. Сначала описан базовый алгоритм, который позволяет анализировать тексты публикаций при помощи тематического классификатора, затем приведена его усовершенствованная версия, учитывающая сетевые связи с помощью эвристического подхода. Исследование методов с применением экспертных оценок и количественных метрик показало, что комбинирование текстовых и сетевых данных значительно повышает точность профилей публикаций. Проверка гипотез о взаимосвязи тематического сходства и сетевой близости публикаций показывает обоснованность предложенного подхода, а также позволяет определить направления дальнейших исследований.

Ключевые слова: сеть публикаций, профиль публикации, теория управления, графовые нейронные сети, анализ текстов.

ВВЕДЕНИЕ

Тематический анализ научных публикаций является важным инструментом для обоснования научных решений и выявления тенденций в различных областях знаний [1–6]. Одним из наиболее распространенных подходов к анализу текстов является тематическое моделирование [7], которое используется для расчета профилей научных публикаций. Однако когда аннотации или тексты публикаций ограничены по объему и (или) содержат неточные термины, использование только текстовой информации может привести к невысокой точности профилей.

Включение в рассмотрение сетевых данных, таких как связи соавторства или цитирования, уже продемонстрировало свою пользу в ряде дисциплин: учет структурных взаимосвязей между публикациями позволяет повысить качество классификации и более адекватно отразить скрытые тематические зависимости [8–10]. В частности, графовые нейронные сети (англ. *graph neural networks*, GNN) [11, 12] зарекомендовали себя как эффективный инструмент для анализа сетей, поскольку они позволяют одновременно учитывать и признаки узлов, и топологию графа.

Целью настоящего исследования является разработка и оценка улучшенных методов построения профилей публикаций, которые комбинируют анализ текстов и сетевой информации. Ниже кратко перечислены основные результаты работы.

- Рассмотрен базовый алгоритм для расчета профилей публикаций на основе тематического классификатора в области теории управления.
- Разработаны «расширенные» алгоритмы, учитывающие сетевые данные. В частности, эвристический метод, который дополняет базовый профиль за счет публикаций, связанных соавторством или цитированием, а также метод на основе графовых нейронных сетей (GNN), позволяющий глубоко интегрировать структурную информацию о связях между публикациями.

• Выполнена оценка эффективности предложенных алгоритмов, показавшая, что их применение дает более высокую точность, чем использование исключительно текстовой информации.

• Исследованы взаимосвязи между тематической схожестью публикаций (оцененной по их профилям) и сетевыми характеристиками (например, общими соседями в графе). В рамках этого анализа были сформулированы и проверены несколько гипотез.

В последующих разделах статьи приводятся детали реализации алгоритмов, используемые метрики и результаты экспериментов.

1. МЕТОДЫ

1.1. Базовый алгоритм расчета профилей

В данной работе для каждой научной публикации l строится *базовый профиль* $p(l)$ с использованием тематического классификатора Информационной системы анализа научной деятельности ИСАНД [13], основанного на принципах, которые изложены в работе [14]. Классификатор представляет собой иерархическую онтологию тем в теории управления (см. фрагмент на рис. 1, на котором длинные названия тем сокращены; полная версия классификатора доступна по адресу https://www.ipu.ru/sites/default/files/page_file/ClassifierCS.xlsx). Разработанная авторами форма[1] использовалась для разметки публикаций группой экспертов в целях дальнейшего сравнения результатов применения базового и сетевого алгоритмов разметки публикаций с экспертной разметкой.

Рис. 1. Тематический классификатор (онтология ИСАНД)

Профиль публикации $p(l)$ — это стохастический вектор $(p_{l1}, p_{l2},\ldots, p_{ln})$, в котором каждая компонента p_{li} равна нормированной частоте терминов, относящихся к теме i в публикации l.

1.2. Расширенные методы расчета профилей

Рассмотрим граф $G(V, E)$, в котором каждая вершина $l \in V$ соответствует одной публикации, а каждое ребро $(l, m) \in E$ означает связь соавторства, т. е. публикации l и m имеют непустое пересечение авторов публикаций $|K(l) \cap K(m)| > 0$, где $K(l)$ – множество авторов (соавторов) публикации l. В исходном графе каждая вершина l инициализируется вектором базового профиля $p(l)$.

1.2.1. Эвристический метод

Для повышения точности базового профиля публикации l учитываются публикации, связанные с l отношением соавторства и вышедшие в свет в течение фиксированного временно́го интервала $\delta \in \mathbb{N}$. По умолчанию $\delta = 4$ года, однако оно может быть скорректировано на основе эмпирических данных. Выбор данного значения обусловлен тем, что в ряде научных дисциплин, включая теорию управления, при оценке научной активности часто рассматривается период в 3–5 лет. В частности, методические рекомендации Высшей аттестационной комиссии (ВАК) предусматривают учет публикаций, вышедших в свет за последние пять лет. Таким образом, выбор $\delta = 4$ года является обоснованным с точки зрения оценки научной активности исследователей.

Расширенный профиль $p_e(l)$ является взвешенной комбинацией

$$p_e(l) = \alpha p(l) + (1-\alpha)\frac{\sum\limits_{m \in L_\delta(l)} w_{lm}\, p(m)}{\sum\limits_{m \in L_\delta(l)} w_{lm}},$$

где $L_\delta(l)$ – множество публикаций, связанных с l и вышедших в свет в пределах временно́го окна δ, а $\alpha \in (0; 1]$ – коэффициент, регулирующий вклад исходного и сетевого профилей. Значение параметра α может быть выбрано эмпирически – например, подобрано по результатам кросс-валидации на отложенной выборке.

Коэффициент $w_{lm} \in [0; 1]$ отражает вклад публикации m в профиль l. В данном случае учитывается доля общих авторов:

$$w_{lm} = \frac{|K(l) \cap K(m)|}{|K(m)|}.$$

[1] URL: https://docs.google.com/forms/d/e/1FAIpQLSfR47ZQyjI9wrMgqRPP85j_uZCeUI95dNFnMR-2ruCfq3XtIg/viewform

Заметим, что, если $\sum_{m\in L_\delta(l)} w_{lm} = 0$, т. е. у l нет связанных публикаций за последние δ лет, профиль $p_e(l)$ по определению равен $p(l)$. В целом эвристический метод позволяет сглаживать «шумы» в базовых профилях, а также формировать сетевой профиль для публикаций с отсутствующей или малоинформативной аннотацией.

1.2.2. Метод на основе графовых нейронных сетей

Для дальнейшего повышения точности профилей применяется графовая нейронная сеть (GNN), обучающаяся на графе $G(V, E)$. Изначально каждый узел $i \in V$ получает вектор признаков $\mathbf{h}_i^{(0)} = p(i)$. На k-м слое GNN вектор $\mathbf{h}_i^{(k)}$ пересчитывается с учетом соседей $\mathcal{N}(i)$ по формуле

$$\mathbf{h}_i^{(k)} = \sigma\left(\sum_{j\in\mathcal{N}(i)} \frac{1}{c_{ij}} W^{(k)} \mathbf{h}_j^{(k-1)} \right),$$

где $\mathbf{h}_i^{(k)}$ – новое представление (профиль) узла i в k-м слое нейронной сети; $W^{(k)}$ – обучаемая матрица весов (параметры модели); σ – нелинейная функция активации (например, ReLU); c_{ij} – нормировочный коэффициент, регулирующий вклад соседних узлов (например, $c_{ij} = \sqrt{\deg(i)\deg(j)}$); $\mathcal{N}(i)$ – множество соседей узла i в графе публикаций.

После прохождения нескольких слоев сети GNN получается итоговый вектор $\mathbf{h}_i^{(K)}$, который можно рассматривать как «глубоко интегрированный» в топологию сети профиль публикации i. При соответствующем обучении (с использованием метрики качества и целевой функции) данный подход позволяет выявить и учесть сложные зависимости между публикациями, что зачастую повышает точность и информативность профилей.

Таким образом, в настоящей работе предложены следующие подходы к расширению базового профиля научных публикаций.

- **Эвристический метод**, задающий линейную комбинацию профиля публикации с профилями связанных работ.
- **GNN-метод**, обеспечивающий более сложное агрегирование признаков на основе обучающей выборки.

Эти подходы обогащают базовые профили, учитывая косвенные тематические связи через соавторство, что способствует решению задач анализа научных публикаций, таких как классификация и рекомендация релевантных статей. Эвристический метод характеризуется высокой интерпретируемостью и простотой реализации; однако его точность может быть ограничена в сложных сетевых структурах, в которых взаимосвязи неочевидны. В свою очередь, метод на основе GNN способен выявлять сложные зависимости между публикациями, что делает его предпочтительным для задач, требующих глубокого анализа структуры соавторства. Однако этот метод может требовать больших объемов данных и значительных вычислительных ресурсов для эффективного обучения модели.

2. ПОСТАНОВКА ЭКСПЕРИМЕНТОВ

Для оценки эффективности методов расчета профилей публикаций в области теории управления использовалась выборка из 20 тысяч статей (база публикаций ИПУ РАН). Данные включали тексты (для построения текстовых признаков) и информацию о соавторстве (для построения сетевых признаков).

Были рассчитаны два вида профилей:

– базовые профили (только по текстам аннотаций),

– расширенные профили (с учетом сетевой структуры, т. е. связей соавторства).

Для количественной оценки качества рассчитанных профилей использовались следующие критерии.

- Экспертные оценки: специалисты предметной области оценивали релевантность тем, приписанных каждой публикации.
- Метрики качества: рассчитывались значения Precision@k, полноты и F_1-меры. Для определения Precision@k было отобрано k наиболее вероятных тем из профиля и выполнено их сопоставление с эталонными темами, определенными экспертами.

Далее в § 3 приведены результаты экспериментов, включающие анализ расстояний между различными профилями и проверку гипотез о взаимосвязи тематического сходства и сетевой близости публикаций.

3. РЕЗУЛЬТАТЫ

Применение расширенного алгоритма, в котором учитывалась сетевая информация, позволило добиться существенного повышения точности относительно базового подхода. Например, эвристический метод (рис. 2) позволил достичь значения Precision@k 37 %, в то время как базовый метод

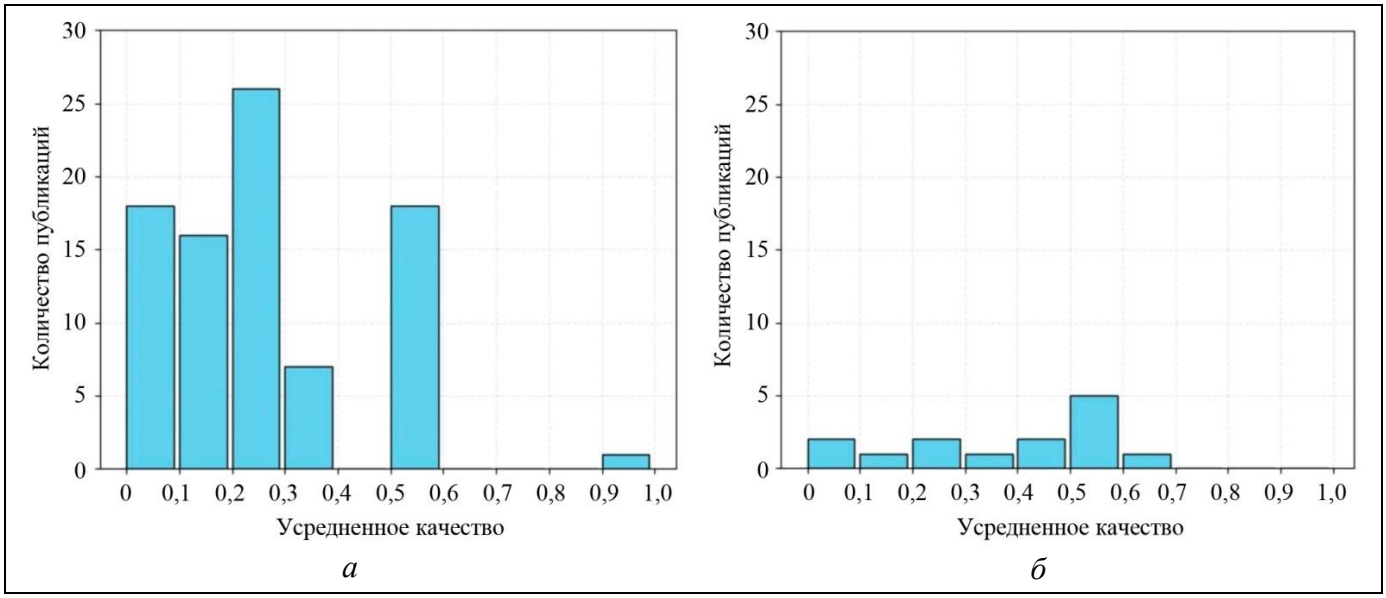

Рис. 2. Оценки качества методов построения профилей публикаций: *а* – базового (BaseAlgo, точность 25 % по аннотациям), *б* – сетевого (HAdvAlgo, точность 37 % по аннотациям). По горизонтали – усредненное качество профилей, по вертикали – число публикаций с таким качеством

обеспечил лишь 25 % (на размеченной экспертами выборке). Это свидетельствует о высокой полезности сетевых признаков для формирования тематических профилей.

В данном исследовании была разработана архитектура графовой нейронной сети (GNN), включающая три последовательных слоя графовой сверточной сети (англ. *graph convolutional network*, GCN) и слои регуляризации Dropout. Для оценки качества модели использовалось разбиение размеченной экспертами выборки (несколько сотен публикаций) в пропорции 70 % / 15 % / 15 % на обучающее, валидационное и тестовое подмножества. Обучение проводилось в течение 100 эпох, при этом итоговые параметры модели выбирались на основе наилучших результатов, достигнутых на валидационной выборке. Среднее значение показателя Precision@k (при k = 3) по всем запускам составило 39 %, это на 2 % выше по сравнению с ранее применявшимся эвристическим методом, что свидетельствует о повышении точности предсказаний.

3.1. Анализ расстояний между профилями

Для более детального сравнения было проанализировано распределение расстояний между расширенными и базовыми профилями для всего множества публикаций (рис. 3). Поскольку использовалась метрика $d \in [0, 1]$ ($L1$), на графике заметны два характерных пика:

- Первый пик при $d = 0$ соответствует случаям, когда у автора только одна публикация. В такой ситуации сетевой профиль практически совпадает с базовым.
- Второй пик при высоких значениях d наблюдается для публикаций, в которых сетевая информация (соавторство) сильно влияет на итоговый профиль и (или) у которых текстовые данные слишком скудны (что затрудняет адекватный расчет базового профиля).

В частности, если аннотация очень коротка или содержит мало релевантных терминов, то базовый профиль может оказаться слабоинформативным. В этих случаях влияние сетевых данных

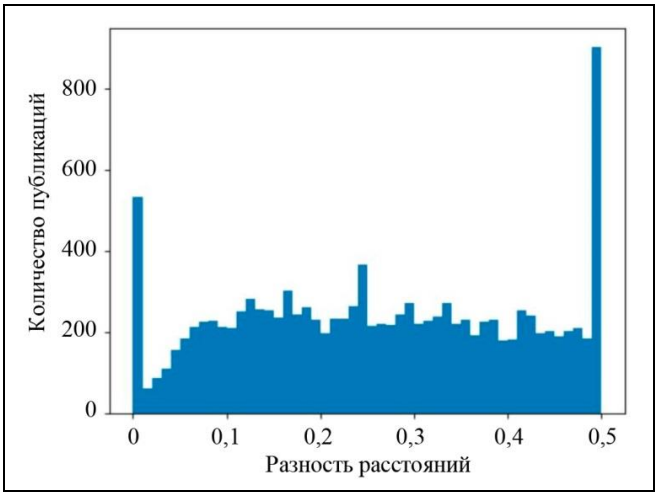

Рис. 3. Распределение расстояний между базовыми и расширенными профилями

оказывается наиболее существенным, и расстояние между профилями (базовым и расширенным) заметно возрастает.

3.2. Проверка гипотез о взаимосвязи тематического сходства и сетевой близости публикаций

Также были исследованы гипотезы, касающиеся связи между тематическим сходством (по профилям) и сетевой близостью (по графу соавторства) публикаций.

Гипотеза 1. *Профили случайно выбранных публикаций отличаются друг от друга.*

Расчеты подтверждают эту гипотезу: среднее расстояние между профилями случайных публикаций составляет около 0,9 (при значениях от 0 до 1), что указывает на значительное разнообразие тем в области теории управления.

Гипотеза 2. *Чем ближе содержание аннотаций двух случайно выбранных публикаций, тем ближе их профили.*

Проверка с помощью векторных представлений (эмбеддингов, англ. *embeddings*; для их построения использовались различные языковые модели – RuSciBert, SciBert, Sentence Embeddings) не показала существенной корреляции. Вероятно, текстовые аннотации слишком коротки или неоднородны, чтобы гарантировать согласование с тематическими профилями, дополненными сетевой информацией. В дальнейшем необходим более детальный анализ природы расхождений (количество терминов, вариативность языка и т. д.)

Гипотеза 3. *Чем больше терминов в аннотации, тем выше корреляция между профилями и векторными представлениями* (*embeddings*) *этих аннотаций.*

Эта гипотеза подтверждается: коэффициент корреляции возрастает от 0,25 (при пяти терминах) до 0,88 (при восьми терминах). Результат подчеркивает важность полноты и точности аннотаций в части используемых терминов.

Гипотеза 4. *Сходство профилей двух публикаций зависит от:*

– наличия связи соавторства,

– совпадения состава авторов.

Данная гипотеза подтверждается: для пар публикаций с общими авторами среднее расстояние между профилями (0,63) оказывается заметно меньше, чем для всех прочих пар (0,88).

Гипотеза 5. *Чем меньше временной интервал между публикациями, тем ближе их профили.*

Предполагается, что в определенные периоды в данной области исследований могут возникать всплески интереса к конкретным технологиям или явлениям (например, «большие данные», «машинное обучение»), что должно быть отражено в содержании публикаций. Такие всплески могут быть связаны с фазами цикла популярности технологий. Однако в области теории управления данная гипотеза не находит подтверждения. Анализ показывает, что для случайных пар публикаций значимых изменений в уровне схожести их профилей в зависимости от временного интервала не наблюдается. В то же время для неслучайных пар публикаций такая зависимость имеет место (неслучайными называются пары публикаций, которые связаны между собой в сети).

Гипотеза 6. *Сходство профилей двух публикаций зависит от топологической силы связи между ними.*

Гипотеза не подтверждена: полученные данные (рис. 4) не выявили значимой корреляции между количеством общих соседей в сети и расстоянием между профилями (расстояние равно нулю, если профили равны). Под топологической силой связи понимается количество связей соавторства между публикациями (топологическая сила связи равна нулю, если связи соавторства отсутствуют).

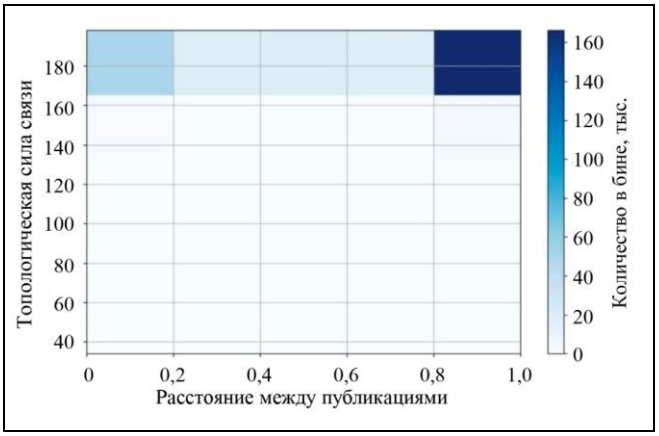

Рис. 4. Зависимость близости публикаций от топологической силы связи между ними. Бин – ячейка, разбиение *N* на *M* равных прямоугольников

4. ОБСУЖДЕНИЕ РЕЗУЛЬТАТОВ

Результаты подтверждают, что учет сетевых признаков (соавторства) действительно повышает точность построения тематических профилей. Эвристический метод ценен своей простотой и интерпретируемостью, что особенно удобно на этапе первичной оценки профилей и анализа предметной области. В то же время применение этого метода показывает и его недостатки: если у авторов мало

статей либо аннотации слишком короткие, то качество базового профиля остается низким, и даже сетевая информация не всегда компенсирует дефицит текстовых данных.

При анализе гипотез было определено значительное разнообразие тем в области теории управления (гипотеза 1) и то, что «текстовое сходство» (гипотеза 2) не обязательно ведет к сходству профилей, особенно при неполных аннотациях. При этом чем больше терминов в публикации, тем больше сходство эмбеддингов определяет сходство профилей (гипотеза 3). Сходство профилей зависит от наличия связи между публикациями (гипотеза 4). Но результаты проверки других гипотез (5 и 6) также говорят о том, что близость по времени или большое число общих соседей еще не гарантируют близости тематических профилей – требуется учет дополнительных факторов и дополнительные исследования.

К ограничениям исследования можно отнести разреженность данных (небольшое число публикаций на одного автора) и неоднородное качество аннотаций, используемых для построения базовых профилей публикаций. Перспективны дальнейшие исследования того, как использование сетей цитирования, ключевых слов, более длинных текстов (полнотекстовые статьи) и продвинутых GNN-моделей (например, с механизмом внимания, англ. *graph attention networks*) может повысить точность профилей.

ЗАКЛЮЧЕНИЕ

Таким образом, предложенные гибридные методы, объединяющие текстовые и сетевые признаки, значительно превосходят базовый (текстовый) подход при построении тематических профилей научных публикаций. Проверка гипотез о тематическом сходстве и сетевой близости показала, что в ряде случаев сетевые связи оказываются гораздо более полезным показателем для определения тематики, чем содержание коротких аннотаций. Полученные результаты будут использованы для развития методов анализа научных публикаций и систематизации знаний в области теории управления.

ЛИТЕРАТУРА

1. *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». – Москва, 2021. – С. 333–338. [*Kryzhanovskaya, S.Y., Vlasov, A.V., Eremeev, M.A., Vorontsov, K.V.* Poluavtomaticheskaya summarizatsiya tematicheskikh podborok nauchnykh publikatsii: zadachi i podkhody // Tezisy dokladov 20-i Vserossiiskoi konferentsii s mezhdunarodnym uchastiem «Matematicheskie metody raspoznavaniya obrazov». – Moscow, 2021. – P. 333–338. (In Russian)]
2. *Shibayama, S., Yin, D., Matsumoto, K.* Measuring Novelty in Science with Word Embedding // PLoS ONE. – 2021. – No. 7. – P. 1–16.
3. *Yuan, W., Liu, P., Neubig, G.* Can We Automate Scientific Reviewing? // Journal of Artificial Intelligence Research. – 2022. – No. 75. – P. 171–212.
4. *Cachola, I., Lo, K., Cohan, A., Weld, D.* TLDR: Extreme Summarization of Scientific Documents // Findings of the Association for Computational Linguistics: EMNLP 2020. – 2020. – P. 4766–4777.
5. *Bao, P., Hong, W., Li, X.* Predicting Paper Acceptance via Interpretable Decision Sets. // In: Companion Proceedings of the Web Conference 2021 (WWW '21). – New York: Association for Computing Machinery, 2021. – P. 461–467.
6. *Kasanishi, T., Isonuma, M., Mori, J., Sakata, I.* SciReviewGen: A Large-scale Dataset for Automatic Literature Review Generation. – arXiv:2305.15186, 2023. – P. 1–19. – DOI: https://doi.org/10.48550/arXiv.2305.15186
7. *Blei, D.M., Ng, A.Y., Jordan, M.I.* Latent Dirichlet Allocation // Journal of Machine Learning Research. – 2003. – No. 3. – P. 993–1022.
8. *Hasegawa, T., Arvidsson, H., Tudzarovski, N.*, et al. Edge-Based Graph Neural Networks for Cell-Graph Modeling and Prediction // Information Processing in Medical Imaging. – 2023. – Vol. 13939. – P. 265–277.
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.
10. *Faber, L., Lu, Y., Wattenhofer, R.* Should Graph Neural Networks Use Features, Edges, Or Both? – arXiv: 2103.06857.arXiv, 2021. – P. 1–12. – DOI: https://doi.org/10.48550/arXiv.2103.06857
11. *Zhou, J., Cui, G., Hu, S.*, et al. Graph Neural Networks: A Review of Methods and Applications // AI Open. – 2020. – Vol. 1. – P. 57–81.
12. *Kipf, T.N., Welling, M.* Semi-Supervised Classification with Graph Convolutional Networks. – arXiv:1609.02907, 2017. – P. 1–14. – DOI: https://doi.org/10.48550/arXiv.1609.02907
13. *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». – Новосибирск, 2023. – С. 89–94. [*Gubanov, D.A., Kuznetsov, O.P., Sukhoverov, V.S., Chkhartishvili, A.G.* O postroenii profilei v tematicheskom prostranstve teorii upravleniya // Materialy 9-i Mezhdunarodnoi konferentsii "Znaniya-Ontologii-Teorii". – Novosibirsk, 2023. – P. 89–94. (In Russian)]
14. *Кузнецов О.П., Суховеров В.С.* Онтологический подход к оценке тематики научного текста // Онтология проектирования. – 2016. – Т. 6, № 1. – С. 55–66. [Kuznetsov, O.P., Sukhoverov, V.S. An Ontological Approach to Determining the Subject Matter of Scientific Text // Ontology of Designing. – 2016. – Vol. 6, no. 1. – P. 55–66. (In Russian)]

Статья представлена к публикации членом редколлегии О.П. Кузнецовым.

Поступила в редакцию 01.11.2024,
после доработки 28.02.2025.
Принята к публикации 06.03.2025.

Губанов Дмитрий Алексеевич – д-р техн. наук, Институт проблем управления им. В. А. Трапезникова РАН, г. Москва
✉ dmitry.a.g@gmail.com
ORCID iD: https://orcid.org/0000-0002-0099-3386

Мельничук Владислав Сергеевич – техник, Институт проблем управления им. В. А. Трапезникова РАН; студент (бакалавр), МГТУ им. Н. Э. Баумана, г. Москва
✉ vs.melnichuk09@gmail.com
ORCID iD: https://orcid.org/0009-0005-8252-0804

CONSTRUCTING SCIENTIFIC PUBLICATION PROFILES BASED ON TEXTS AND COAUTHORSHIP CONNECTIONS (IN THE FIELD OF CONTROL THEORY AND ITS APPLICATIONS)

D. A. Gubanov* and V. S. Melnichuk**

*,***Trapeznikov Institute of Control Sciences, Russian Academy of Sciences, Moscow, Russia,
**Bauman Moscow State Technical University, Moscow, Russia

*✉ dmitry.a.g@gmail.com, **✉ vs.melnichuk09@gmail.com

Abstract. The calculation of scientific publication profiles is crucial in the systematization of scientific knowledge and support for scientific decision-making. This paper proposes a method for forming publication profiles in the field of control theory, based on the integration of text analysis and coauthorship network analysis. We describe a basic algorithm that analyzes publication texts by a thematic classifier as well as its enhanced version that considers network connections within a heuristic approach. The methods are examined using expert assessments and quantitative metrics; according to the examination results, combining textual and network data significantly improves the accuracy of publication profiles. Hypotheses about a relationship between the thematic similarity and network proximity of publications are tested, and the approach proposed is validated accordingly. In addition, directions for further research are identified.

Keywords: publication network, publication profile, control theory, graph neural networks, text analysis.