УДК 519.17:004.8

АЛГОРИТМИЧЕСКАЯ СИСТЕМА ИЗВЛЕЧЕНИЯ И СТРУКТУРНОГО АНАЛИЗА НАУЧНЫХ ПУБЛИКАЦИЙ

Прохоров Матвей Алексеевич[1]
Научный руководитель – доктор биол. наук, профессор Кубряк О.В.[2]
Научный консультант – канд. техн. наук, доцент Ковальчук С.В.[1]
[1]Университет ИТМО
[2]НИУ «МЭИ»

Введение

Современный стремительный рост научного корпуса текстов порождает методологические трудности при реконструкции генезиса и трансформации научных понятий. Показательным примером такой трансформации является концепция «стресса», введенная Гансом Селье, впоследствии трансформировавшаяся в междисциплинарное понятие. Проследить эту эволюцию валидными методами непросто. Существующие подходы в основном основаны на традиционных библиометрических показателях (количество публикаций, цитирование, совместное цитирование) или методах тематического моделирования, которые идентифицируют скрытые темы [1], однако оба подхода ограничены в возможности формализовать структурные связи между понятиями и моделировать их динамику. Развитие сетевого подхода к анализу науки позволило визуализировать и исследовать библиометрические сети [2], тем не менее такие методы преимущественно фиксируют срез на определённый момент, но плохо исследуют динамику и редко интегрируются с вероятностными моделями тематических состояний.

В последние годы предпринимаются попытки объединить тематическое моделирование с векторными представлениями документов в пространстве, что позволяет учитывать семантическую близость текстов и частично моделировать динамику тематических состояний [3]. Тем не менее сохраняется методологическая фрагментарность анализа: вероятностные модели тем, сетевые методы и алгоритмы структурной кластеризации чаще применяются раздельно и не образуют единой вычислительной системы. Таким образом, актуальной задачей является разработка алгоритмической системы, которая сочетает в себе автоматическое извлечение публикаций, графовую структуризацию научных связей и анализ их динамики во времени, что позволит повысить воспроизводимость и формализованность исследований эволюции научных концептов.

Основная часть

Предлагаемое решение состоит в разработке модульной алгоритмической системы, которая будет представлять научные статьи о стрессе в виде формализованной модели. Эта система автоматически осуществляет поиск публикаций в научной базе данных PubMed, выполняет их тематический отбор и приводит к единому структурированному формату. Далее статьи рассматриваются как элементы единой модели, в которой выделяются ключевые сущности – документы, авторы и основные термины, а также связи между ними, отражающие терминологические, авторские и временные отношения. В результате формируется сеть, описывающая внутреннюю организацию исследуемой научной области.

В отличие от традиционных способов, преимущественно основанных на подсчёте публикаций и цитирований, предлагаемый подход рассматривает научное знание как совокупность взаимосвязанных объектов, допускающих количественную оценку структурных характеристик: плотности связей, центральности элементов, устойчивости кластеров и характер их группировки. Последовательный анализ изменений сети во

времени позволяет проследить развитие научной идеи и выявить переходы между различными состояниями её структурной организации. Такой подход обеспечивает более строгую формализацию эволюции научного понятия и позволяет выявлять закономерности его трансформации.

Оптимальность решения достигается за счёт модульной архитектуры и инкрементальной обработки данных, что обеспечивает масштабируемость и воспроизводимость результатов. Оригинальность подхода заключается в интеграции процедур тематической верификации и сетевого анализа в единую вычислительную схему, где каждый этап обработки данных подчинён общей модели структурного представления знания. Это позволяет не только количественно оценивать текущее состояние исследуемой области, но и выявлять скрытые структурные сдвиги, возникающие при формировании новых направлений исследований. Разработанная система может быть адаптирована к анализу других научных концептов, выступая универсальным инструментом математически ориентированного исследования эволюции научного знания.

Выводы

Результаты исследования могут быть использованы как инструмент структурного анализа научных направлений: разработанная алгоритмическая система позволяет автоматически формировать модель исследуемой области, выявлять ключевые темы, значимые публикации и авторов, фиксировать появление новых направлений и отслеживать их развитие во времени, что делает её применимой при подготовке аналитических обзоров, стратегическом планировании исследований и оценке междисциплинарных связей.

Литература

1. Blei D. M., Ng A. Y., Jordan M. I. Latent Dirichlet Allocation // Journal of Machine Learning Research. – 2003. – Vol. 3. – P. 993–1022.
2. van Eck N. J., Waltman L. Visualizing bibliometric networks // Measuring Scholarly Impact / eds. Y. Ding, R. Rousseau, D. Wolfram. – Cham : Springer, 2014. – P. 285–320. – DOI: 10.1007/978-3-319-10377-8_13.
3. Dieng A. B., Ruiz F. J. R., Blei D. M. Topic modeling in embedding spaces // Transactions of the Association for Computational Linguistics. – 2020. – Vol. 8. – P. 439–453.