

АНАЛИЗ ПУБЛИКАЦИОННОЙ АКТИВНОСТИ НАУЧНЫХ КОЛЛЕКТИВОВ

Негбоев А.Н.

Белорусский государственный университет информатики и радиоэлектроники
г. Минск, Республика Беларусь

Шульдова С.Г. – канд. техн. наук, доцент

Данное исследование посвящено анализу метода интеллектуального анализа данных публикационной активности научных коллективов. Предложен комплексный подход, объединяющий кластеризацию, семантический анализ текстов и технологии оперативного анализа данных (OLAP) для оценки продуктивности сотрудников и выявления тематической направленности исследований. Практическая реализация выполнена на основе данных сотрудников научного коллектива за период 2013–2021 годов. Результаты анализа демонстрируют динамику роста числа публикаций, структуру цитируемости и ключевые направления научной деятельности. Показано, что интеграция методов повышает точность оценки и выявляет скрытые закономерности, что ценно для управления научными коллективами.

Современная наука характеризуется экспоненциальным ростом объема публикаций и связанной с ними информации, что делает традиционные методы анализа публикационной активности недостаточными для получения глубоких инсайтов. Подсчет числа публикаций, расчет индекса Хирша или анализ цитируемости [1], хотя и остаются важными инструментами, не позволяют в полной мере оценить качественные аспекты научной деятельности, такие как тематическая релевантность, междисциплинарные связи или влияние на развитие дисциплин. Публикационная активность является ключевым индикатором научной продуктивности, отражающим как индивидуальный вклад исследователей, так и эффективность работы коллективов [2]. Однако в условиях глобальной конкуренции и роста требований к наукометрическим показателям возникает необходимость в более сложных подходах, которые выходят за рамки простого измерения количественных метрик. Интеллектуальный анализ данных, включающий методы машинного обучения, семантический анализ и многомерное моделирование, позволяет не только описать текущую ситуацию, но и прогнозировать тенденции, а также поддерживать стратегическое планирование научной деятельности.

Целью данной работы является разработка и апробация методов интеллектуального анализа публикационной активности научного коллектива. В качестве объекта исследования выбраны данные о публикациях сотрудников за период 2013–2021 годов, что позволяет проследить динамику и выявить ключевые направления исследований.

Для анализа публикационной активности предложен комплексный подход, объединяющий три основных метода: оперативный анализ данных с использованием OLAP-куба, кластеризацию методом k-средних и семантический анализ текстов. Каждый из методов решает свою задачу, дополняя общую картину.

Технологии оперативного анализа данных (Online Analytical Processing, OLAP) применены для изучения временной динамики и структуры публикационного потока. Разработан многомерный OLAP-куб, реализованный с использованием СУБД Microsoft SQL Server и службы MS SQL Server Analysis Services (SSAS). Куб включает группы мер (количество публикаций, цитируемость, число соавторов) и измерения (издатель, автор, публикация, год). Связь «многие-ко-многим» между авторами и публикациями трансформирована в гиперкуб, где таблица фактов содержит количественные показатели, а таблицы измерений описывают атрибуты, такие как тип публикации и ученая степень автора. Процесс извлечения, очистки и загрузки данных (ETL) организован для обработки данных из открытых источников, включая устранение дубликатов и заполнение пропусков.

Для группировки сотрудников по уровню публикационной активности и тематической направленности использован алгоритм k-средних, который минимизирует внутригрупповую дисперсию. Пусть $X = \{x_1, x_2, \dots, x_n\}$ – набор данных, где $x_i = (P_i, C_i)$ – вектор, включающий число публикаций P_i и цитирований C_i сотрудника (i). Целевая функция минимизации задается как:

$$J = \sum_{j=1}^k \sum_{x_i \in s_j} \|x_i - \mu_j\|^2, \quad (1)$$

где μ_j – центроид (j)-го кластера, s_j – множество объектов в кластере, (k) – число кластеров (в данном случае $k = 3$).

Евклидово расстояние между объектами рассчитывается как:

$$d(x, \mu_i) = \sqrt{\sum_{j=1}^m (x_j - \mu_{ij})^2}, \quad (2)$$

где (m) – размерность вектора признаков.

Алгоритм реализован с использованием библиотеки scikit-learn на языке Python, что обеспечило высокую вычислительную эффективность.

Для выявления тематической направленности применен контент-анализ, основанный на частотности употребления ключевых слов и их совместной встречаемости. Процесс включал этапы: извлечение текстов из публикаций, очистку данных (удаление стоп-слов, нормализация), подсчет частоты слов и кластеризацию ключевых слов методом k-средних. Анализ проводился на корпусе из 150 публикаций, доступных в открытом доступе.

Практическая реализация предложенного подхода выполнена на данных о публикациях сотрудников научного коллектива за период 2013–2021 годов. Общий объем обработанных записей составил более 600 публикаций, включая статьи, монографии и тезисы конференций. Результаты анализа можно разделить на три ключевые группы.

OLAP-анализ выявил устойчивый рост числа публикаций на 30% за указанный период, с пиком в 2017–2018 годах (среднегодовой прирост – около 4%). Этот всплеск может быть связан с профессиональными достижениями сотрудников, такими как присвоение ученых званий или участие в крупных проектах. Структура потока показала преобладание журнальных статей (60%) над тезисами конференций (25%) и монографиями (15%).

Средняя цитируемость на публикацию составила 12 ссылок, а средний индекс Хирша сотрудников – 4, с максимальным значением 8 у лидеров коллектива. Кластеризация методом k-средних выделила три группы сотрудников: «высокоактивные» (свыше 20 публикаций и 100 цитирований), «средние» (10–20 публикаций) и «низко активные» (менее 10 публикаций). Доля высоко цитируемых работ (HCR) составила 15%, что указывает на наличие значимых исследований.

Семантический анализ определил три основных направления исследований:

- 1) численные методы и оптимизация (40% текстов, ключевые слова: «алгоритм», «оптимизация»);
- 2) машинное обучение и анализ данных (30%, ключевые слова: «моделирование», «нейросети»);
- 3) разработка программного обеспечения (20%, ключевые слова: «программирование», «системы»).

Временной анализ показал рост интереса к машинному обучению после 2018 года (увеличение числа публикаций на 25%), что соответствует глобальным трендам. Однако междисциплинарные темы составили менее 10% корпуса, что указывает на узкую специализацию коллектива. Анализ выявил сильные стороны коллектива: стабильный рост продуктивности и активное внутреннее сотрудничество (кластеры соавторства охватывают 70% сотрудников). Среди слабостей – низкая внешняя цитируемость (менее 20% ссылок от авторов вне коллектива) и ограниченный доступ к полным текстам (50% публикаций), что снизило точность семантического анализа.

Интеграция методов интеллектуального анализа данных доказала свою эффективность в оценке публикационной активности научного коллектива. OLAP-куб обеспечил гибкость анализа структуры потока, кластеризация выявила группы сотрудников по уровню продуктивности, а семантический анализ определил приоритетные направления исследований. Полученные результаты имеют прикладное значение: они могут быть использованы для оптимизации публикационной стратегии, повышения видимости работ в международном сообществе и диверсификации тематики исследований. Однако подход имеет ограничения, связанные с качеством и полнотой исходных данных. Например, отсутствие доступа к полным текстам части публикаций повлияло на точность тематического анализа. Для повышения эффективности требуется дальнейшая автоматизация процессов сбора и обработки данных, а также интеграция дополнительных источников, таких как Scopus или Web of Science [3]. Перспективы исследования включают разработку прогнозных моделей на основе временных рядов и расширение анализа на междисциплинарные связи, что может усилить вклад коллектива в мировую науку.

Список использованных источников:

1. Хирш Х.Э. Индекс для измерения научной продуктивности // *Proceedings of the National Academy of Sciences*. 2005. № 102(46). Р. 16569–16572.
2. Тронин В.Г. Оценка результатов научно-исследовательской работы и наукометрия : учебное пособие / В.Г. Тронин, А.Р. Сафиуллин. – Ульяновск : УлГТУ, 2019. – 136 с.
3. Симоненко Т.В. Наукометрия: методология, инструменты, практическое применение : монография / Т.В. Симоненко. – Минск: Беларуская навука, 2018. – 343с.