

МЕТОДИКА СОПОСТАВЛЕНИЯ ОЦЕНОК ВАЖНОСТИ НАУЧНЫХ РЕЗУЛЬТАТОВ В ПРОФЕССИОНАЛЬНОМ СООБЩЕСТВЕ И ШИРОКОЙ ОБЩЕСТВЕННОСТИ

METHODS OF SCIENTIFIC IMPORTANCE COMPARISON BETWEEN PROFESSIONAL AND PUBLIC SOCIETIES

Широков Никита Константинович – младший научный сотрудник, Государственная публичная научно-техническая библиотека СО РАН (Российская Федерация), e-mail: shirokov@gpntbsib.ru

Костюк Даниил Максимович – младший научный сотрудник, Государственная публичная научно-техническая библиотека СО РАН (Российская Федерация), e-mail: kostuk@gpntbsib.ru,

Shirokov Nikita Konstantinovich – junior researcher, The State Public Science and Technology Library, e-mail: shirokov@gpntbsib.ru

Kostuk Daniil Maksimovich – junior researcher, The State Public Science and Technology Library, e-mail: kostuk@gpntbsib.ru

Аннотация: В рамках одного из проектов фундаментальных научных исследований ГПНТБ СО РАН возникла задача сопоставления оценки научных результатов профессиональным сообществом и широкой общественностью. В качестве предмета оценки были выбраны совокупные показатели цитируемости научных направлений в СМИ (охват, количество перепечаток, упоминание в соцсетях), полученные с помощью тематического моделирования и методов машинного анализа данных новостей научно-информационного ресурса «Новости сибирской науки».

Abstract: In the scope of the project of fundamental scientific research of the State Public Scientific Technical Library of the SB RAS, the task of comparing the assessment of scientific results by the professional community and the general public arose. The aggregate indicators of citation of scientific directions in the media (coverage, number of reprints, mentions in social networks), obtained using thematic modeling and methods of machine analysis of news data of the scientific information resource «sib-science.info», were selected as the subject of the assessment.

Ключевые слова: анализ данных, тематическое моделирование, альтметрики, наукометрия.

Keywords: data analysis, thematic modeling, altmetrics, scientometrics

Анализ новостного медиаконтента стал важной частью измерения влияния результатов научных исследований на общество в альтметриках – альтернативных метриках импакта публикаций на выходящую за пределы профессионального сообщества среду. Однако в классическом варианте альтметрики (сервисы Altmetrics, Plum Analytics) измеряют показатели конкретных научных статей и опираются преимущественно на англоязычные

ресурсы, что снижает их значимость для российского общества. В Российской Федерации важным элементом научной политики является оценка результативности научных организаций. С учётом приоритетов Стратегии научно-технологического развития и других программных документов, определение влияния отдельных результатов научных исследований и научной деятельности организаций в целом является важной задачей.

В рамках одного из проектов фундаментальных научных исследований ГПНТБ СО РАН планируется расширение научно-информационного ресурса «Новости сибирской науки» на российскую науку в целом. Ресурс «Новости сибирской науки» был создан в 2015 году для представления сообщений из печатных и электронных СМИ и на данный момент содержит более 45 000 новостных сообщений. В качестве подзадачи стоит создание аналитического блока, который бы позволил получить характеристику представления в СМИ учёных и научных организаций. Особенно интересным является в этом свете сопоставление оценки научных результатов профессиональным сообществом и широкой общественностью. Поэтому одной из целей создания блока стала разработка методологии сопоставления наукометрической и альтметрической оценки значимости выбранных научных направлений. В качестве предмета оценки была выбрана значимость научных тематик.

Для проведения такого анализа необходимо было решить следующие задачи:

1. Выбрать перспективные и значимые научные тематики для анализа.
2. С помощью наукометрических инструментов получить оценку важности выбранных научных направлений.
3. Разработать и реализовать механизм классификации новостных сообщений по научным направлениям.
4. Получить оценку значимости научных направлений на основе данных СМИ (охват, количество перепечаток, упоминание в социальных сетях и т. д.)
5. Выполнить сопоставление наукометрической и альтметрической оценки значимости выбранных научных направлений.

Множество существующих в настоящий момент методов классификации текстов базируется на терминологической близости. Текст представляется в виде вектора в евклидовом пространстве, где оси координат – это термы, n-граммы или лексемы, выделяемые из текста, а координатой по оси является статистическая информация о них. Таким образом текст может быть представлен в виде частотных векторов встречаемости слов на основе схем BOW (Bag Of Words), TF (Term Frequency), TF*IDF, TF*CHI и других. Однако для эффективной работы алгоритмов классификации, основанных на частотных векторах, требуется предобработка текстов, целью которой является уменьшение размера словаря. Предобработка включает в себя различные этапы, такие как удаление неинформативных слов, очистка пунктуации и т. д. Согласно данным, приведенным в тексте диссертации на тему «Иерархическая классификация коллекций документов» за авторством А.А. Кузьмина, «предобработка позволяет улучшить качество классификации

документов для некоторых языков в 10-50 раз, а также уменьшить размер словаря на 50%» [1]. Новости, представленные на ресурсе «Новости сибирской науки», имеют формат HTML, в связи с чем была произведена их очистка от HTML-тегов и спецсимволов. Полученный текст был приведён к нижнему регистру, а затем обработан с помощью алгоритма исправления опечаток и орфографии YandexSpeller. Оставшиеся слова были приведены к нормальной форме с лемматизации — процесса приведения словоформы к лемме (её нормальной словарной форме). В отличие от более простых методов и алгоритмов (стемминг, отсечение), лемматизация преобразует слова к словарной форме, что упрощает для человека дальнейший анализ результатов классификации. Для удаления неинформативных слов, не несущих смысла для классификации, а также незначимых частей речи, таких как союзы и предлоги, были использованы словари стоп-слов. При предобработке и очистке текста важно учитывать порядок применяемых процедур [2]. В нашей работе, к примеру, процесс лемматизации был выполнен до удаления стоп-слов, так как при её применении могли бы заново появиться стоп-слова.

Для классификации текста по темам наш выбор был сделан в пользу модели LDA. В данном алгоритме важным параметром является количество тем. При малом количестве тем классификация может привести к крупным кластерам, содержащим различные аспекты. В то время как слишком большое количество тем приводит к появлению схожих мелких кластеров, различить которые будет затруднительно [3]. При классификации текстов новостей для выбора количества тем был использован итеративный подход с учётом количества тематик классификатора Scopus второго уровня, количества основных тем ресурса и соответствующих им тегов, содержащихся в самих новостях. В связи с этим было запущено сразу несколько моделей-кандидатов с различным количеством тем, которые впоследствии сравнивались на предмет различий и интерпретируемости. При количестве в 40 кластеров был получен удовлетворительный результат, тем не менее содержащий как несколько крупных кластеров с общими терминами, так и несколько мелких кластеров, содержащих 0,3-0,5 % от всех токенов.

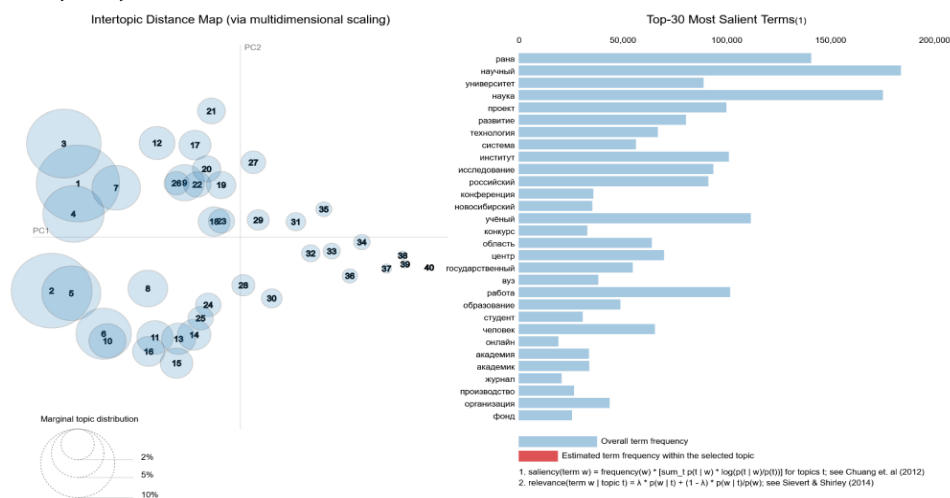
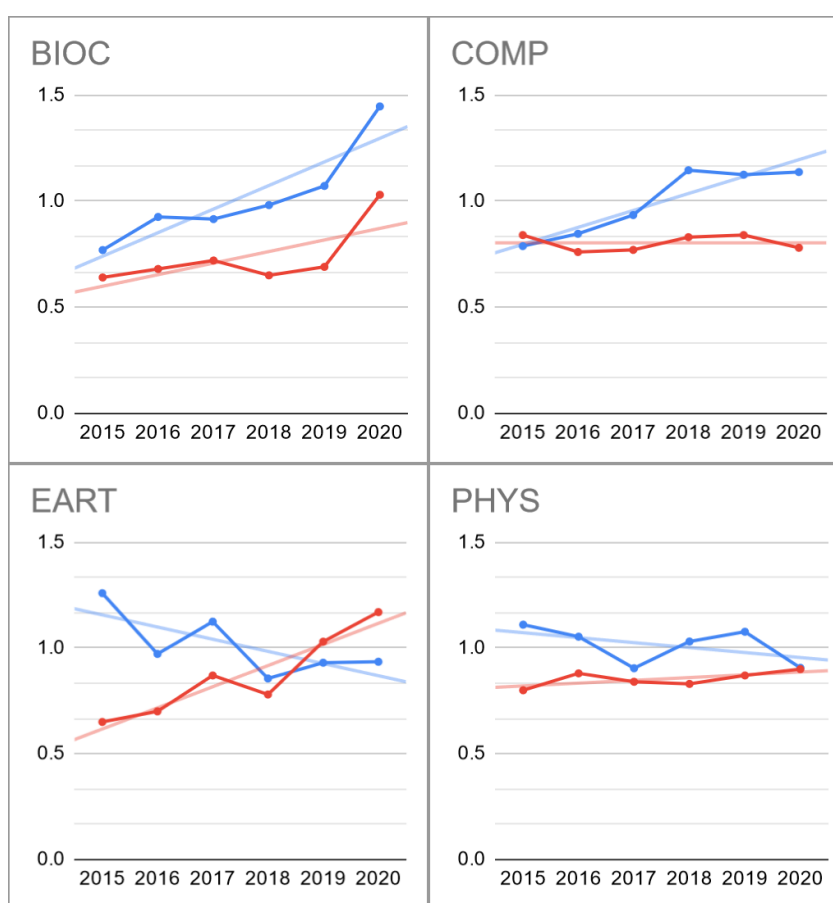


Рис. 1. Визуализация результата работы модели LDA на 40 кластерах

На следующем этапе были проведены интерпретация и сопоставление полученных кластеров с классификатором Scopus. В результате сопоставления 16 из 40 кластеров удалось соотнести с десятью тематиками (62 % от всех новостных сообщений портала). 14 из 40 кластеров были определены как не относящиеся напрямую к научным тематикам (научная политика, внутренние новости институтов, новости-анонсы и т.д.). Оставшиеся 10 кластеров были определены в смежную группу, так как содержали различные научные термины, не относящиеся к одной тематике. Результатом классификации новостного сообщения в модели LDA стал вектор вероятностей принадлежности текста к полученным кластерам.



Красный — данные FWCI из Scopus / SciVal
Синий — данные FWMI из ресурса «Новости сибирской науки»

Рис. 2. Сравнение FWCI (научное сообщество) и FWMI (широкая общественность)

Для десяти полученных тематик Scopus с помощью платформы SciVal были взяты показатели наукометрической оценки FWCI (Field-Weighted Citation Impact). Данная метрика показывает соотношение общего числа фактически полученных цитирований и числа цитирований, ожидаемого (среднего) для предметной области. С помощью неё можно оценить важность результатов научного направления в профессиональном сообществе. Для получения оценки важности научного направления среди широкой общественности

в SciVal существует метрика FWMM (Field-Weighted Mass Media), показывающая соотношение упоминаний в СМИ к ожидаемому среднемировому показателю для предметной области, типа и года публикации. Эта метрика позволяет оценить преобладание или, наоборот, малозначительность упоминаний научного направления в средствах массовой информации. При её расчёте учитываются просмотры в онлайн и печатных изданиях, а также упоминания в соцсетях, блогах и комментариях. Однако в SciVal показатели метрики FWMM отсутствуют, в связи с чем был предложен и построен аналог метрики FWMM – FWMI (Field-Weighted Media Importance).

При подсчёте суммарного числа просмотров тематики в год, просмотры каждой новости были взяты с весом, соответствующим вероятности попадания данной новости в рассматриваемую тематику. Также взвешивалось и число статей внутри тематики. Далее для каждого года и тематики вычислялся показатель FWMI по формуле: $(V(T, Y) / \sum V(T, Y)) / (V(T) / \sum V(T))$, где $V(T, Y)$ – просмотры тематики за год, $\sum V(T)$ – суммарные просмотры тематики за весь период, а суммирование ведётся по всем тематикам. Результаты сравнения FWCI и FWMI для некоторых тематик приведены на рис. 2 (BIOC – Биология и биологические науки, COMP – Информатика и компьютерные науки, EART – науки о земле, PHYS – Физика и астрономия).

СПИСОК ЛИТЕРАТУРЫ

1. Кузьмин, А. А. Иерархическая классификация коллекций документов : дис. ... канд. физ.-мат. наук : 05.13.17 / Арсентий Александрович Кузьмин. – М., 2017. – 120 л.
2. Grimmer, J. A Bayesian Hierarchical Topic Model for Political Texts: Measuring Expressed Agendas in Senate Press Releases / J. Grimmer // Political Analysis. – 2010. – Vol. 18. – P. 1–35.
3. Applying LDA Topic Modeling in Communication Research: Toward a Valid and Reliable Methodology / Daniel Maier [et al.] // Communication Methods and Measures. – 2018. – Vol. 12, iss. 2/3. – P. 93–118.
4. SciVal Metric: Field-Weighted Citation Impact (FWCI) [Electronic resource] // SciVal Support Center. – Mode of access: https://service.elsevier.com/app/answers/detail/a_id/28192/supporthub/scival/p/10961/28192/. – Date of access: 10.05.2021.