

СОДЕРЖАНИЕ

Кун С., Чжан Ц., Чжан Д., Бу И., Дин И., Ся Ф. Геном успеха в науке	3
Фахимифар С., Горби А., Ауслоос М. Мы стоим на ненадежных плечах? Влияние цитирования отозванных из публикации статей на предыдущие и последующие опубликованные статьи: исследование базы данных Web of Science	19

РЕДАКЦИОННАЯ КОЛЛЕГИЯ

ГЛАВНЫЙ РЕДАКТОР

д.филол.н. ГИЛЯРЕВСКИЙ Р.С.

ЧЛЕНЫ РЕДКОЛЛЕГИИ:

**к.т.н. БЫКОВ В.А. (РОССИЯ), к.физ.-мат.н. ВАРДАНЯН Г. Г. (АРМЕНИЯ),
д.т.н., проф. ВОЙТОВ И. В. (БЕЛАРУСЬ), МИРАЛИЕВ К. Х. (ТАДЖИКИСТАН),
МОЛДОШЕВА Д. А. (КЫРГЫЗИЯ)**

РЕДАКТОРЫ:

ОВЧЕНКОВА Е.А., САРКИСЯН Д.Б.

Геном успеха в науке*

СЯНЦЗЕ КУН
(Xiangjie KONG)

Чжэцзянский технологический
университет, Китай

Цзюнь ЧЖАН
(Jun ZHANG)

Даляньский технологический
университет, Китай

Да ЧЖАН
(Da ZHANG)

Университет Майами, США

И БУ
(Yi BU)

Пекинский университет, Китай

ИН ДИН
(Ying DING)

Техасский университет в Остине, США

ФЭН СЯ
(Feng XIA)

Федеральный университет Австралии,
Австралия

ВВЕДЕНИЕ

Прогресс нашего общества в значительной степени связан с неустанными комментариями исследований учеными. Чтобы признать и поддержать

*В этом документе описывается, как выявлять и оценивать причинно-следственные факторы для улучшения научного влияния. В настоящее время анализ научного влияния может быть полезен для различных видов научной деятельности, включая заявку на финансирование, рекомендации наставников, а также поиск потенциальных коллег и т.д. Общепризнано, что высококвалифицированные ученые часто имеют больше возможностей получить награды в качестве поощрения за свою усердную работу. Поэтому ученые тратят огромные усилия на создание научных достижений и улучшение научного влияния в течение своей научной жизни. Однако каковы определяющие факторы, которые контролируют академический успех ученого? Ответ на этот вопрос может помочь ученым более эффективно проводить свои исследования. Под таким углом в нашей статье представлены и проанализированы причинно-следственные факторы, которые имеют решающее значение для академического успеха ученых. Сначала мы предлагаем пять основных факторов ориентированных на: статью, автора, место проведения, учреждение, и факторы времени. Затем мы применяем новейшие передовые алгоритмы машинного обучения и метод складного ножа для оценки важности каждого причинно-следственного фактора. Наши эмпирические результаты показывают, что факторы, ориентированные на автора и статьи, имеют наибольшее значение для будущего успеха ученых в области компьютерных наук. Кроме того, мы обнаруживаем интересный феномен: *h*-индекс ученых в одном и том же учреждении или университете на самом деле очень близки друг к другу.*

их вклад, этим выдающимся исследователям предоставляется ряд наград и возможностей для проведения исследований. В последнее время многие исследователи сосредоточены на том, как определить выдающихся ученых и как стать успешным ученым, а также предлагается множество показателей с разных аспектов для количественной оценки этого научного успеха [19, 25]. Однако в числе этих разнообразных показателей еще предстоит выяснить, какой фактор (факторы) является решающим и спо-

* Перевод Kong X., Zhang J., Zhang D., Bu Y., Ding Y., Xia F. The gene of scientific success // ACM Trans. Knowl. Discov. Data. — 2020. — <https://arxiv.org/pdf/2202.08461.pdf>

способствует научному успеху. Более того, изучение причинно-следственных связей является фундаментальной проблемой в машинном обучении с применениями в различных областях, таких как биология, экономика, эпидемиология и информатика [31, 38]. Инспирированная приведенными выше наблюдениями, наша статья сосредоточена на изучении причинно-следственных связей, которые играют жизненно важную роль в академическом успехе ученых.

Количественная оценка научного успеха ученых всегда была интересной темой, привлекающей к своему изучению исследователей с разным опытом [13, 30, 37, 40]. Цель науки о достижении успеха состоит в том, чтобы сначала понять лежащий в основе механизм, а затем найти генеративную модель для прогнозирования того, какие ценности научного успеха будут приняты во внимание вне причинно-следственных факторов. На основе количества цитирований был предложен ряд оценочных показателей, таких как импакт-фактор журнала [14], g-индекс [12], h-индекс [16] и т.д. Б. Ван Хаутен [34] определяет научное влияние как экспертную оценку коллегами исследований, академических работ и других достижений, важность научного влияния зависит от того, насколько их исследовательские достижения ценятся, признаются и цитируются другими. Для измерения научного влияния исследователи определили различные контролирующие факторы, чтобы отразить ряд особенностей субъектов науки. В числе этих факторов подсчет цитирований рассматривался как основной фактор для оценки научного влияния из-за его простоты и эффективности.

Помимо метрик, основанных на цитировании, ученые также исследуют этот вопрос с точки зрения сетевых топологий. Изначально алгоритмы ранжирования по важности, такие как PageRank [27] и HITS [17], предназначены для ранжирования важности веб-страниц. В последнее время, вдохновленные этими алгоритмами ранжирования по важности, исследователи также широко используют их для оценки научного влияния в академических сетях [2, 11]. Учитывая достоинства как алгоритмов PageRank, так и алгоритмов HITS, Ван и др. [36] предлагают метод MRCoRank для измерения влияния научных субъектов в гетерогенных академических сетях путем взаимной системы поощрений.

Более того, развитие социальных сетей позволяет ученым регулярно обмениваться своими статьями в Twitter или Facebook. Этот эффект обмена информацией больше не ограничивается академическими социальными сетями. Он распространился на многие цифровые библиотеки и широко используется на платформах социальных сетей. Наряду с этой тенденцией, альтметрия предлагается в качестве еще одного критерия для измерения популярности ученых и их публикаций путем оценки полученного ими общественного внимания. Исследователи, тем временем, начинают использовать альтметрию для

количественной оценки научного влияния ученого, поскольку она может быстро фиксировать раннее влияние. Например, Борнманн и др. [6] нормализуют количество публикаций в Twitter, чтобы измерить влияние исследований, а затем используют его для межобластных сравнений. Притом, многие методы исследования изучают корреляцию между альтметрией и методами, основанными на подсчете цитирования, путем статистического анализа их взаимосвязи [9, 39].

Оценка научного влияния может пролить свет на различные практические вопросы, такие как заявки на получение премий или финансирования, трудоустройство и выбор консультанта [24]. Как правило, успешные ученые получают дополнительную возможность приобретать ресурсы для исследований, получать гранты и распространять свои исследовательские идеи более широко. Поэтому ученые стремятся постоянно повышать свое научное влияние. Однако какие факторы имеют причинно-следственную связь с научным успехом? А именно, какие факторы наиболее подходят для оценки научного успеха? Вышеуказанные вопросы до сих пор остаются нерешенными. Поэтому в этой статье мы сначала проводим исследования, выявляя причинно-следственные факторы, которые способствуют научному успеху, а затем оцениваем важность причинно-следственных связей этих факторов. Мы берем наиболее часто используемый индекс h в качестве показателя для оценки влияния ученых и выявляем причинно-следственные факторы, которые приводят к высокому индексу h у ученых.

Из-за проблем с конфиденциальностью и технологических ограничений доступ к информации о публикациях гораздо проще предоставить по сравнению с другими источниками данных. Она (информация) также может в значительной степени отражать академические способности или вклад соответствующих ученых. Как правило, название, ключевые слова, авторы, учреждения, место проведения, страницы, и опубликованные даты издания могут быть получены напрямую. На основе этих данных можно извлечь и рассчитать множество импакт-факторов, как это делается в большинстве текущих работ. Хотя, в отличие от предыдущей работы, наш метод не фокусируется на улучшении показателей оценки, а стремится на выявление причинно-следственных факторов, влияющих на академические успехи ученых по их публикациям. Чтобы ответить на этот вопрос, мы разделили факторы влияния на несколько категорий, как показано на рис. 1. Это фактор статьи, фактор автора, фактор места проведения, фактор учреждения и фактор времени. Для каждого фактора мы предлагаем конкретные и интуитивно понятные показатели, отражающие академические качества каждого ученого. После этого, используя алгоритмы машинного обучения и метод складного ножа, мы исследуем влияние каждого фактора на академические успехи ученых.



Рис. 1. Иллюстрация факторов, воздействующих на влияние ученого

Вклад. Наше исследование в основном сосредоточено на выявлении причинно-следственных факторов академической эффективности ученых. В целом, в этой статье мы вносим следующий вклад:

- *Новые функции.* Мы представляем пять потенциальных причинно-следственных факторов с учетом нового коэффициента Джини для институтов.
- *Выявление причинно-следственных связей.* Используя алгоритмы машинного обучения и метод складного ножа, мы обнаружили, что факторы, ориентированные на автора и статьи, сильно коррелируют с их академическими успехами.
- *Новое понимание.* Наши результаты предоставляют исследователям новый и эффективный метод повышения их научного влияния.

Статья организована следующим образом. После раздела **ВВЕДЕНИЕ** обсуждается связанная работа. Далее в разделе **ИДЕНТИФИКАЦИЯ ПРИЧИННО-СЛЕДСТВЕННЫХ ФАКТОРОВ** указаны предлагаемые факторы научного влияния. В разделе проверяются причинно-следственные факторы, оценивается их значимость, и они подтверждаются на большом наборе данных об ученых. Затем мы завершаем нашу работу разделом **ЗАКЛЮЧЕНИЕ И ДАЛЬНЕЙШАЯ РАБОТА.**

СВЯЗАННАЯ РАБОТА

Научное влияние изучалось на протяжении десятилетий исследователями из самых разных дисциплин. В течение долгого времени подсчет цитирования широко применялся для измерения научного влияния. Наряду с этой тенденцией исследователи предложили различные показатели,

основанные на цитировании. С развитием академических сетей ученые также рассматривают проблему научного влияния с точки зрения важности сети. Эти показатели могут быть использованы как для оценки влияния, так и для прогнозирования будущих успехов в науке. В этом разделе мы представим связанную работу по вышеупомянутым аспектам.

Подсчет цитирований впервые был использован для количественной оценки влияния журналов. С тех пор исследователи предложили множество основанных на цитировании методов для измерения научного влияния [33], таких как h-индекс [16] и g-индекс [12]. Некоторые ученые утверждают, что цитирования не следует рассматривать как равнозначные [5]. Еще один пример, который следует упомянуть здесь: и А, и В цитируют С. Ранее цитаты из А и В считались одинаковыми, но, если А - из высокоцитируемой статьи, а В - нет, цитирования следует различать. Это похоже на идеи, связанные с алгоритмом PageRank. Был предложен ряд подходов для определения важности цитирования [33]. Все эти методы применяли подсчет цитирования как важную часть метрики оценки, но все они вносят некоторые улучшения, так как простая опора на подсчет цитирования является неадекватной для оценки влияния [35].

Помимо использования цитирования для количественной оценки научного влияния, научные сети в настоящее время часто применяются для изучения подобных проблем, поскольку сети содержат различные типы субъектов и отношений [23]. Алгоритмы PageRank и HTS широко используются

для измерения научного влияния в академических сетях. На основе этих двух алгоритмов был предложен ряд сетевых оценочных показателей [42]. Учитывая влияние различных академических сетевых структур, ученые применяют модифицированные алгоритмы ранжирования по важности для оценки влияния различных научных субъектов [3]. Кроме рассмотрения сетевых топологий, некоторые исследователи также обнаруживают новые функции и взаимосвязи для оценки научного влияния. Ван и др. [36] оценивают влияние научных субъектов, исследуя особенности текста в разнородных академических сетях. Из-за развивающегося характера академических сетей некоторые исследования также учитывают динамику цитирования и появление новых объектов или отношений для оценки научного влияния [4, 43].

В дополнении к использованию основанных на цитировании и сети возможностей для оценки научного влияния, ученые также пытаются изучить соответствующие факторы, которые очень важны для будущей академической успеваемости, и предсказать будущее влияние [15, 26]. Ван и др. [35] проверяют эффективность ранних цитирований в прогнозировании потенциальных цитирований статей [7, 18]. Стегехьюс и др. [32] используют два важных фактора, а именно историческую информацию о цитировании и импакт-фактор журнала, для прогнозирования распределения цитирования статей.

Прогнозирование потенциального влияния ученых, h -индекс и предстоящие цитирования – все это входит в сферу прогнозирования будущего влияния [8, 35]. Акуна и др. [1] используют количество статей, h -индекс и академический возраст ученого, чтобы предсказать его влияние. Метод линейной регрессии используется для прогнозирования будущего влияния выдающихся ученых в области математики, физики и биологии. И эти авторы обнаружили, что академический возраст ученых на самом деле играет значительную роль в прогнозировании научного влияния [22]. Кроме того, другие авторы [10] изучили вопрос о том, какая статья может увеличить h -индекс ученого с помощью метода линейной регрессии. Они обнаружили, что в числе шести факторов тема и место проведения имеют очень важное значение для прогнозов.

Прогнозирование научных последствий с помощью причинно-следственных связей – это недавно появившаяся область исследований. В отличие от предыдущих методов [1, 8, 35], методы, основанные на причинно-следственных выводах, сначала выявляют потенциальные причинные факторы, а затем используют их для прогнозирования научного влияния. Более того, предыдущие методы рассматривали только одну сторону для оценки научного влияния, пренебрегая анализом и ранжированием важности каждого причинного фактора.

ИДЕНТИФИКАЦИЯ ПРИЧИННО-СЛЕДСТВЕННЫХ ФАКТОРОВ

Исследователи десятилетиями изучали проблему научного влияния и предлагают различные факторы влияния. Однако до сих пор не существует формального определения научного влияния и общепринятого стандарта его оценки. Какие из этих ранее изученных импакт-факторов наиболее важны для академического успеха ученых? Получение ответа на этот вопрос может помочь исследователям проводить свои исследования более эффективно. В этом разделе мы представим несколько новых факторов влияния, упорядочим существующие факторы и классифицируем их по различным категориям.

Факторы, ориентированные на статьи

Как правило, большинство предыдущих работ предпочитают использовать подсчет цитирований и число статей для количественной оценки научного влияния. Хотя помимо этих двух показателей существуют различные факторы, основанные на статьях, которые влияют на динамику научного влияния. Чтобы определить репрезентативные возможности статей, мы сначала анализируем элементы, связанные с факторами, ориентированными на статьи.

Подсчет цитирований ($Cits$) и число публикаций (Num_{pub}) являются основой факторов, основанных на статьях. Средние ссылки для каждого ученого ($Ave_{ci}^{a_i}$), самые высокие ссылки ($Hi_{ci}^{a_i}$), самые низкие ссылки ($Lo_{ci}^{a_i}$) могут быть получены непосредственно через значения их общего количества ($Cits$) и (Num_{pub}). Более того, качество статьи зависит не только от ее содержания, но и от популярности темы. Например, ранее широкий спектр данных не мог быть собран и обработан из-за технических ограничений. В то время как с развитием технологий обработки данных и продвижением эры больших данных все больше внимания уделяется документам, связанным с тематикой больших данных. Следовательно, темы статьи также могут касаться ее влияния. Чтобы это учесть, мы предлагаем показатель степень популярности темы статьи (article's topic popular degree- ATP), которая может быть рассчитана в соответствии со следующим уравнением.

$$ATP(p_i) = \frac{\sum_{w=1}^m Num(w)^{p_i}}{\sum_{i=1}^n Num(i)} \quad , \quad (1)$$

где (p_i) представляет статью, w -ключевое слово статьи p_i , $Num(w)$ - количество w , m - количество ключевых слов в статье p_i , $Num(i)$ - количество ключевых слов в статьях и (n) - общее количество публикаций.

Помимо вышеупомянутых факторов, основанных на цитировании, при измерении научного влияния также необходимо учитывать качество источников ссылок. Как правило, у каждого ученого есть список публикаций, и каждая публикация содержит ряд источников ссылок. Цитаты могут рассматриваться как академические признания от других исследователей. Точно так же авторы статьи также получают информацию из ее источников ссылок. Таким образом, источники ссылок могут повлиять на качество статьи. Прежде всего, наибольшее (Hi_{ci}^{ref}), среднее ($Ave_{ci}^{a_i}$), наименьшее (Lo_{ci}^{ref}) количество цитирований и среднее количество источников (Ave_{num}^{ref}) являются наиболее прямыми измерениями для количественной оценки качества источников. Помимо цитирования, влияние источников ссылок также используется для оценки влияния ссылок, поскольку многие исследователи склонны цитировать статьи из источников с высоким влиянием, независимо от релевантности между статьями.

Чтобы измерить релевантность между статьями (Rel_{ref}), мы сначала решаем эту проблему с точки зрения авторов. В соответствии с публикациями каждого автора, области их исследований могут быть представлены через ключевые слова конкретных статей. Поэтому мы используем различия между ключевыми словами авторов для расчета релевантности между статьями и источниками. Для ее количественной оценки применяется информационная энтропия, и формула расчета выглядит следующим образом:

$$Rel_{ref}^{p \rightarrow q} = - \sum_{i=1}^r W_i \log_2 (W_i) , \quad (2)$$

где ($Rel_{ref}^{p \rightarrow q}$) представляет релевантность между статьей q и ее источником ссылки p , (W_i) - частота слов в статье q и p - информация о ключевых словах в p , а r - общее количество слов.

Кроме того, необходимо также учитывать взаимосвязь между статьями и их источниками. Из-за отсутствия полных текстов статей мы используем косинусное сходство для измерения релевантности между статьями и названиями и ключевыми словами источников и ссылок на них. Для каждой статьи и ее источника мы извлекаем последовательность слов ($m_1, m_2, m_3, \dots, m_n$) из их названий и ключевых слов. Затем вектор может быть получен на основе приведенной выше последовательности для каждой статьи. В соответствии с этими векторами релевантность между статьей и ее ссылкой может быть рассчитана следующим образом:

$$Sim(p_1, p_2) = \frac{\sum_{i=1}^n (V_{p_1,i} * V_{p_2,i})}{\sqrt{\sum_{i=1}^n V_{p_1,i}^2} * \sqrt{\sum_{i=1}^n V_{p_2,i}^2}} , \quad (3)$$

где $Sim(p_1, p_2)$ представляет собой соответствие между статьями p_1 и p_2 , V_{p_1} является вектором p_1 и V_{p_2} - вектором p_2 .

Факторы, связанные с местом проведения

Помимо показателя, основанного на цитировании, PageRank также может использоваться для измерения качества места проведения, что отражает научный успех автора. Чтобы оценить важность мест проведения, сначала вычисляются значения PageRank ($PR(v_i)$) в сети места проведения. Затем среднее количество цитирований статей, опубликованных в местах проведения ($Ave_{ci}^{v_i}$), используется для измерения качества мест проведения. Кроме того, с помощью концепции h -индекса ученого мы вычисляем h -индекс места проведения ($h(v_i)$). В частности, определение h -индекса места проведения похоже на обычный порядок расчета h -индекса, и значение h -индекса места проведения равняется h , у которого, по крайней мере, h статей в месте проведения имеет h ссылок

$$PR(v_i) = \sum_{j=1}^n Ave_{cj}^{v_i} . \quad (4)$$

Факторы, ориентированные на автора

Кроме факторов, ориентированных на статьи, факторы, представляющие признаки ученых, также жизненно важны для их влияния. h -индекс каждого ученого (h_{ai}) и значение PageRank (PR_{ai}) в сети сотрудничества (коллоборации) являются интуитивными факторами, указывающими на влияние ученого. Между тем, импакт-фактор журнала (JIF) может быть рассчитан на их основе и широко применяется для измерения влияния журналов из-за его простоты. В соответствии с концепцией JIF предлагается авторский импакт-фактор (AIF). Аналогично, AIF ученого в год t - это AIF ученого (Ave_{ci}) за Δt годы до года t . Помимо подсчета цитирований, сумма оценок PageRank статей ученых в сети цитирования (PR_{pub}) также может указывать на их важность.

В дополнение к этим двум факторам ученые предложили несколько хорошо известных факторов для количественной оценки динамики влияния ученых. Значение Q широко применяется для выявления процесса взаимного усиления влияния ученых на их статьи [30] и остается стабильным на протяжении всей академической карьеры ученых. Формула расчета значения Q выглядит следующим образом:

$$Q(a_i) = e^{\langle \log_{cia} \rangle} - \mu_p , \quad (5)$$

где $Q(a_i)$ представляет значение Q ученого, $\langle log_{c_{ia}} \rangle$ - среднее количество цитирований a_i в логарифмическом порядке, (α) - статья α от автора a_i и (μ_p) - среднее потенциальное влияние статей.

На протяжении своей карьеры каждый ученый взаимодействует с множеством исследователей из разных дисциплин. Ученые получают выгоду от академических обменов и дискуссий с другими исследователями, а также улучшают свое собственное научное влияние. Следовательно, способности соавторов также могут повлиять на качество их статей и влияние ученых. Чтобы определить влияние соавторов, предлагается несколько факторов. Как правило, h -индекс соавторов отражает их способности. Основываясь на этом, можно легко получить ряд факторов. Максимальные ($h \max_{ci}^{a_i}$) и средние значения ($have_{co}^{a_i}$) соавторов ученых могут быть получены непосредственно через h -индекс каждого автора. Затем мы используем различия между h_{a_i} -индексом и $h \max_{co}^{a_i} (hdif_{a_i})$, чтобы представить расстояние между влиятельными соавторами и (a_i).

Затем мы рассматриваем влияние разнообразного исследовательского опыта соавторов на влияние ученых. Поскольку взаимодействия между исследователями становятся все более частыми, интеграция ученых из разных дисциплин также оказывает положительное влияние на продвижение достижений науки и технологий. Чтобы измерить диапазон дисциплин соавторов, мы применяем теорию энтропии. Подробную информацию о конкретных дисциплинах и институтах ученых можно получить из набора данных. Для каждого ученого мы количественно оцениваем разнообразие его соавторов ($Div(a_i)$), используя теорию энтропии. Разнообразие вычисляется в соответствии со следующими уравнениями:

$$Div(a_i)_{inst} = - \sum_{m=1}^r w_m \log_2 (w_m) \quad (6)$$

$$Div(a_i)_{key} = - \sum_{p=1}^q k_p \log_2 (k_p) \quad (7)$$

$$Div(a_i) = Div(a_i)_{inst} + Div(a_i)_{key} \quad (8)$$

где $(Div(a_i)_{inst})$ и $(Div(a_i)_{key})$ представляют разнообразие a_i ученых из институтов коллег автора и ключевые слова их статей для автора a_i ($Div(a_i)$) указывает на разнообразие a_i всех коллег. (w_m) - частота слова m в общей информации учреждений коллег a_i и общее количество слов m в уравнении (9). (k_p) - частота слова p во всех ключевых словах статей коллег a_i и общее количество слов p .

Институционально-ориентированные факторы

Воздействие институтов на влияние ученых также необходимо учитывать, поскольку вопросы финансирования исследований или политики могут существенно повлиять на прогресс исследователей в их исследованиях. Академические достижения ученых также могут зависеть от способностей их коллег, поскольку они часто могут обмениваться исследовательскими идеями и методами. Как правило, мы исследуем влияние институтов с двух основных точек зрения: академическая среда ученых и экономические факторы.

Мы оцениваем академическую среду с точки зрения коллег. При проведении исследования люди, как правило, обмениваются идеями со своими соавторами или коллегами. Кроме того, исследователи также подвержены влиянию влиятельной группы или отдельных лиц в их учреждении. Это влияние обычно определяется как давление со стороны коллег (или социальное давление). Принимая во внимание такое давление, мы пытаемся определить значение давления со стороны коллег на академическую производительность ученых. Другими словами, существует ли между ними реальная связь? Чтобы ответить на указанные вопросы, мы предложили несколько факторов, чтобы выявить корреляцию между академическими успехами ученых и их коллегами. Первоначально мы оцениваем исследовательские способности коллег ученых. Для каждого ученого по набору данных можно рассчитать h -индекс его коллеги (h_{col}), количество публикаций (Num_{pub}^{col}), количество ссылок ($Cits_{col}$) и рейтинг PageRank (PR_{col}).

Кроме того, мы используем концепцию коэффициента Джини из экономической области для описания академической репутации учреждения. В коэффициенте Джини первоначально используется определение глобальных кривых Лоренца для вычисления распределения доходов в экономической области. Его значение колеблется от 0 до 1. Чем больше значение, тем больше экономическое неравенство. В нашей статье мы количественно оцениваем коэффициент Джини учреждений, используя значения h -индекса ученого, цитирования и количества статей. Коэффициент Джини учреждения может быть рассчитан следующим образом:

$$G(i) = 1 - \frac{1}{n} \left(2 \sum_{m=1}^{n-1} P_m + 1 \right), \quad (9)$$

где $G(i)$ представляет значение коэффициента Джини учреждения i и n представляет собой количество исследовательских групп внутри учреждения i . Для исследовательской группы m указывается групповой индекс среди групп. Кроме того, (P_m) - это доля суммы групп m во всех значениях учре-

ждения i . Следовательно, в соответствии со значениями h -индекса ученого, цитирований и количества статей, репутация каждого учреждения рассчитывается с использованием трех значений коэффициента Джини, которыми являются $G(i)^h$, $G(i)^{Cit}$ и $G(i)^{pub}$.

Факторы времени

Предыдущие исследования подтвердили влияние временной динамики на научное влияние, например, на предсказание восходящих звезд в академических кругах. Что касается молодых исследователей, то у них может быть этап быстрого роста после начала академической карьеры. Производительность в этот период очень важна для их будущих академических успехов. Мы предлагаем два фактора времени, чтобы отразить это явление. Первый из них - академический возраст (N_{im}^{years}), который представляет собой годы с момента публикации учеными своих первых научных работ. Другим фактором является динамика h -индекса ученых в течение Δt лет. В этой статье мы устанавливаем $\Delta t = 3, 5, 7$, а затем вычисляем разницу ($Hindex-dif$) между прогнозируемым временем и Δt годами в прошлом.

$$\rho = \frac{cov(X, Y)}{\sigma_X \sigma_Y} = \frac{E(XY) - E(X)E(Y)}{\sqrt{E(X^2) - E^2(X)}\sqrt{E(Y^2) - E^2(Y)}}, \quad (10)$$

где cov - ковариация между двумя группами результатов и указывает на их стандартное отклонение. Его значение колеблется от -1 до 1, а корреляция ва-

рируется от самой отрицательной до самой положительной.

Исходя из вышеупомянутых факторов, мы стараемся как можно полнее перечислить актуальные показатели академической эффективности ученых. Эти показатели подразделяются на пять основных категорий: факторы, ориентированные на статью, факторы, ориентированные на автора, факторы, ориентированные на место проведения, факторы, ориентированные на учреждение, и факторы времени. Эти факторы описаны в табл. 1. Между тем, мы в первую очередь исследуем корреляцию между этими факторами и h -индексом ученого самым прямым способом. Коэффициент корреляции Пирсона применяется для измерения релевантности (соотношения) между двумя результатами ранжирования. Процедура расчета коэффициента корреляции Пирсона показана следующим образом.

Согласно табл. 1, мы видим, что факторы, ориентированные на автора и статью, являются наиболее коррелирующими факторами в числе всех предложенных показателей, за которыми следуют факторы, ориентированные на время. Чтобы дополнительно представить явление линейной корреляции между h -индексом и другими факторами, мы затем приводим результаты четырех наиболее значимых факторов. Как показано на рис. 2, количество цитирований и публикаций ученых, академические годы и различия h -индекса в Δt годах сильно коррелируют с будущим h -индексом ученых. С увеличением академического возраста индекс h продолжает увеличиваться до достижения академического возраста 15 лет, после чего остается стабильным.

Таблица 1

Описания причинно-следственных факторов и корреляции

	Функция	Описание	Корреляция
Статья	$Cits$	Количество цитирований ученых.	0,7629
	Num_{pub}	Количество публикаций ученых.	0,7782
	Ave_{ci}	Средняя цитируемость каждого ученого.	0,2772
	Hi_{ci}	Самая высокая цитируемость каждого ученого.	0,2349
	Lo_{ci}	Наименьшая цитируемость каждого ученого.	0,2067
	ATP	Степень популярности темы статьи.	0,0134
	Hi_{ci}^{ref}	Наибольшее количество ссылок.	0,1648
	Ave_{ci}^{ref}	Среднее количество ссылок.	0,1439
	Lo_{ci}^{ref}	Наименьшее количество ссылок.	0,0648
	Ave_{num}^{ref}	Среднее количество ссылок.	0,2496
	Rel_{ref}	Релевантность между статьями.	0,0174
	$Sim(p_1, p_2)$	Косинусное сходство между статьями.	0,1437
Место проведения	$PR(vi)$	Значения PageRank места проведения в сети места нахождения статьи	0,2146
	Ave_{ci}^{vi}	Среднее количество цитирований статей, опубликованных в их местах проведения.	0,1924
	$h(vi)$	h -индекс места проведения	0,2081

	Функция	Описание	Корреляция
Автор	$h(a_i)$	Значение h -индекса каждого ученого.	0,9782
	PRa_i	Значение PageRank каждого ученого в сети соавторов.	0,6274
	AIF	Импакт-фактор автора.	0,3826
	$Qvalue$	Значение Q автора.	0,5394
	$hmax_{co}^{a_i}$	Максимальное значение h -индекса соавторов ученого.	0,8253
	$Num_{co}^{a_i}$	Количество соавторов ученого.	0,426
	$have_{co}^{a_i}$	Среднее значение h -индекса соавторов ученого.	0,482
	$hlo_{co}^{a_i}$	Наименьшее значение h -индекса соавторов ученого.	0,275
	$hdif_{a_i}$	Разница между максимальными и наименьшими значениями h -индекса соавторов ученого.	0,538
	$Div(a_i)$	Разнообразие соавторов.	0,1743
Учреждение	h_{col}	h -индекс коллег ученых.	0,2947
	Num_{pub}^{col}	Количество публикаций коллеги ученого.	0,1368
	$Cits_{col}$	Количество цитирований коллеги ученого.	0,1937
	PR_{col}	Оценка PageRank коллеги ученого.	0,0264
	$G(i)^h$	Коэффициент Джини по h -индексу учреждения.	0,0937
	$G(i)^{Cit}$	Коэффициент Джини по количеству цитирований учреждения.	0,0153
	$G(i)^{pub}$	Коэффициент Джини по количеству публикаций учреждения	0,1632
	GNP	Величина ВВП страны учреждения.	0,1937
Фактор времени	Num_{years}	Академический возраст ученого	0,5683
	$Hindex-dif$	Разница между h -индексом ученого и Δ годы в прошлом.	0,6248

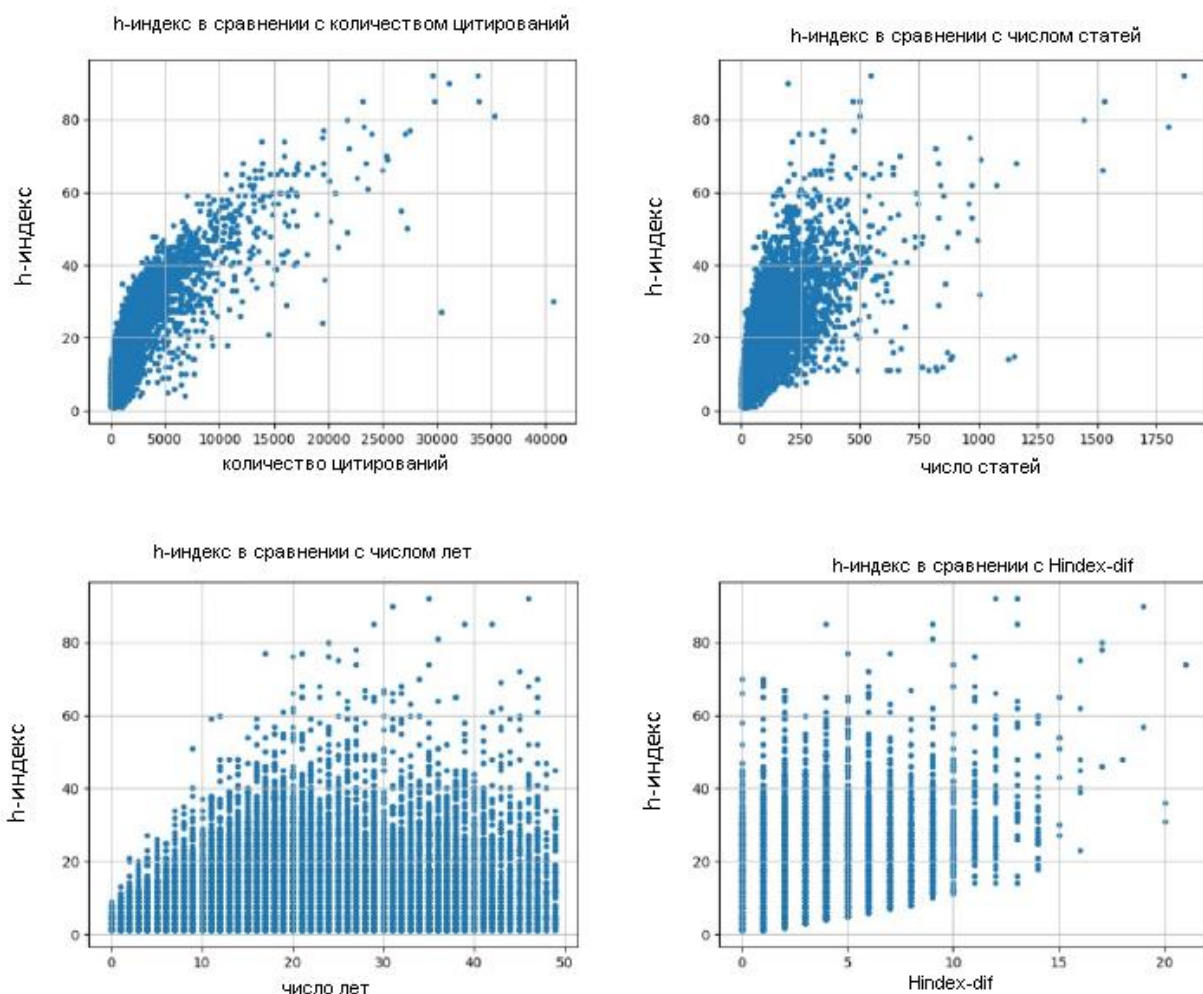


Рис. 2. Явления линейной регрессии между h -индексом и четырьмя наиболее значимыми факторами

Нетрудно понять высокое значение ($Cites$, Num_{pub} и h_a). В то время как факторы, ориентированные на место проведения, и факторы, ориентированные на учреждение, по-видимому, отрицательно коррелируют с h -индексом ученого. Однако эта таблица не может точно отразить эффективность этих факторов в прогнозировании будущего академического успеха ученых, поскольку эти факторы могут вместе выявлять одно и то же явление. Из этой таблицы можно увидеть только линейную корреляцию между ними, и их эффективность в прогнозировании h -индекса ученого анализируется в следующем разделе.

Проверка причинно-следственных факторов

В этом разделе мы исследуем влияние вышеупомянутых факторов на прогнозирование h -индекса ученого. Чтобы исследовать их производительность, мы используем передовые методы машинного обучения. В их числе мы применяем XGBoost, линейную регрессию, деревья решений с градиентным бустингом, а также дерево классификации и регрессии по отдельности на наборе данных. Затем, сравнивая производительность, мы находим наиболее подходящий метод машинного обучения.

Построение структурированной причинно-следственной модели

В основе структурной теории причинности лежит "модель структурной причинности (SCM) [21, 29, 41]". Поэтому в нашей статье мы сначала представляем простую причинно-следственную модель, как показано на рис. 3. Здесь S - коллаيدر: стрелки «сталкиваются» в точке S . Путь $A \rightarrow S \leftarrow B$ заблокирован. Другими словами, A не связан с B через S . С учетом коллайдера S , причинные факторы независимы друг от друга. Мы называем эту структуру V-структурой [20]. В этой V-структуре A и B являются потомками (родительскими элементами) для S . Однако в нашей статье научный успех может определяться и зависеть от множества причинно-следственных факторов. Поэтому мы расширяем рис. 3(а) до гетерогенной структурированной причинно-следственной модели, как показано на рис. 3(б), сформулированной в виде следующего уравнения (рис. 4).

$$P(S) = P(S|F_1) \dots P(S|F_n) \quad (11)$$

Обнаружение причинного фактора

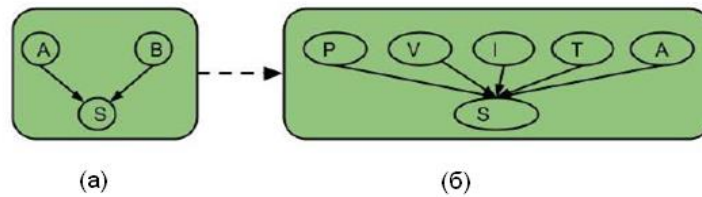


Рис. 3. Модель причинно-следственного вывода V-образной структуры

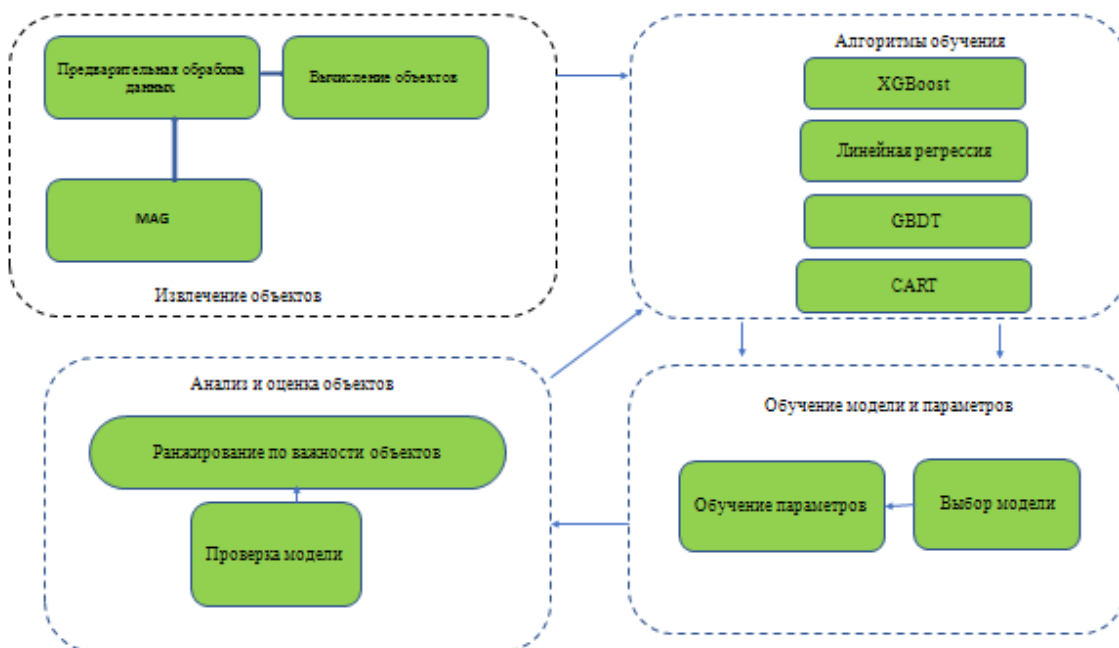


Рис. 4. Структура причинно-следственного вывода

Модель SCM, состоящая из двух наборов переменных, X и Y , и набора функций F , которые определяют, как присваиваются значения каждой переменной ($X_i \in X$). Здесь мы предполагаем, что, учитывая результат прогнозирования, функция представляет влияние как функцию прямых причин и предельных потерь с параметрами обучения (θ_1). После того, как мы определим факторы, способствующие нашему научному успеху, нам нужно измерить значимость для каждого причинного фактора. Что касается большого научного набора данных наблюдений, то мы применяем передовые методы машинного обучения, чтобы выявить причинно-следственные связи и то, как различные причинно-следственные факторы, включая автора, статью, место проведения, учреждение и время, способствуют пониманию научного успеха.

$$S = F(X, Y, \epsilon; \theta_1) \quad (12)$$

Потеря функции

В этом разделе мы применяем следующие четыре передовых метода машинного обучения для оценки причинно-следственных связей.

XGBoost: XGBoost – это масштабируемая непрерывная система бустинга деревьев, которая работает быстрее, чем самые современные широко используемые методы. Идея алгоритма абустинга состоит в том, чтобы объединить множество слабых классификаторов вместе, чтобы сформировать сильный классификатор, а XGBoost – это поддерживающая модель дерева, которая объединяет множество моделей дерева регрессии CART для формирования сильного классификатора. Ее дерево бустинга в основном состоит из двух частей, которые представляют собой объект изучения регуляризации и процесс градиентного бустинга дерева.

Линейная регрессия (LR): Регрессионный анализ широко используется для прогнозирования и предсказания, а также может быть использован для определения того, какие из всех независимых переменных связаны с зависимой переменной. Линейная регрессия требует, чтобы модель была линейной по параметрам регрессии. Функция прогнозирования используется для моделирования данных, и данные могут быть использованы для оценки неизвестных параметров. Линейная регрессия быстра в моделировании и запускается в случае больших объемов данных.

Деревья решений с градиентным бустингом (GBDT): GBDT – это итеративный алгоритм дерева решений, который включает в себя множество деревьев решений, а конечный результат равен сумме решений всех деревьев. Суть GBDT заключается в том, что каждое дерево усваивает остаток от суммы всех предыдущих выводов дерева, кото-

рый представляет собой сумму реальных значений после добавления прогнозируемых значений. Он может обнаруживать множество отличительных особенностей и их комбинаций.

Деревья классификации и регрессии (CART):

Его можно использовать для создания дерева классификации или дерева регрессии. Когда CART используется в качестве дерева классификации, атрибуты объектов могут быть непрерывными или дискретными, а дерево классификации CART использует коэффициент Джини при разделении узлов. Когда CART используется в качестве дерева регрессии, атрибуты наблюдения должны быть непрерывного типа. Поскольку метод наименьшего абсолютного отклонения (LAD) или наименьшего квадратного отклонения (LSD) обычно используется при выборе атрибутов объектов путем разделения узлов, атрибуты объектов также имеют непрерывный тип. В нашей статье мы применяем его в качестве дерева регрессии для прогнозирования будущего влияния ученого на основе входных переменных.

Набор данных

В этой статье мы используем два набора данных из разных дисциплин. Одним из них является вспомогательный набор данных, извлеченный из Microsoft Academic Graph (MAG). Набор данных MAG содержит подробную информацию о документе, включая название, ключевые слова, авторов, учреждения, места проведения, дату публикации и цитаты из 27 макрообластей и 306 подобластей. Весь набор данных включает более 35 миллионов статей, 38 миллионов авторов и свыше 324 миллионов взаимосвязей ссылок. Мы используем вспомогательный набор данных, включающий 79321 профиль ученых и 105123 статьи, посвященные области компьютерных наук от ученых с завершенной академической карьерой.

Другой является подмножеством Американского физического общества (APS). Набор данных APS содержит информацию о статье по физике с указанием названия, авторов, учреждений, мест проведения, даты публикации и цитирования. Весь набор данных включает 540232 статьи, 394801 автора и более 6 миллионов взаимосвязей ссылок на 12 журналов APS. Мы используем вспомогательный набор данных, включающий статьи PRC и PRE, 80360 профилей ученых и в общей сложности 98011 статей.

Показатели оценки

Чтобы оценить производительность различных алгоритмов и факторов обучения, мы используем четыре типичных показателя, включая MAE (Средняя абсолютная ошибка), MAPE (Средняя абсолютная процентная ошибка), MSE (Среднеквадратичная ошибка), ACC (Точность) и R^2 . Учитывая истинное

значение y и прогностическое значение $\hat{y}$, вышеупомянутые оценочные показатели могут быть рассчитаны следующим образом:

$$MAE = \frac{1}{n} \sum_{i=1}^n |\hat{y}_i - y_i| \quad (13)$$

$$MAPE = \frac{100}{n} \sum_{i=1}^n \left| \frac{y_i - \hat{y}_i}{y_i} \right| \quad (14)$$

$$MSE = \frac{1}{n} \sum_{i=1}^n |\hat{y}_i - y_i|^2 \quad (15)$$

$$ACC = \frac{1}{n} \sum_{i=1}^n I(f(y_i) = \hat{y}_i) \quad (16)$$

$$R^2 = 1 - \frac{\sum_{i=1}^n |y_i - \hat{y}_i|^2}{\sum_{i=1}^n |y_i - \bar{y}|^2} \quad (17)$$

Проверка с помощью экспериментов

С помощью алгоритмов обучения и факторов, которые мы представили выше, можно предсказать h -индекс ученого. Мы используем информацию за предыдущие Δt годы для обучения и реальные данные за 2015 год (MAG) или 2013 год (APS) для проверки. На тренировочном наборе мы выполняем 5-кратную перекрестную проверку, чтобы настроить гиперпараметр моделей. Для XGBoost и GBDT основные гиперпараметры, которые мы настроили, включают скорость обучения, максимальную глубину деревьев, частоту под-выборок, частоту под-признаков и коэффициент регуляризации. Для CART основные гиперпараметры, которые мы настроили, включают скорость обучения, максимальную глубину деревьев. Для LR основные гиперпараметры, которые мы настроили, включа-

ют скорость обучения и коэффициент регуляризации. Все гиперпараметры настраиваются с помощью поиска по сетке в пространстве параметров. Результаты проиллюстрированы с точки зрения MAE, MAPE, MSE, ACC и (R^2).

В табл. 2 показаны прогностическая эффективность различных методов для показателей оценки, упомянутых выше, в наборе данных MAG. MSE, MAE и MAPE используются для сравнения результатов прогнозирования и истинных значений. В таблице указывается корреляция между прогнозируемыми результатами и истинными значениями, а также указывается точность. Следовательно, по их значениям можно сделать вывод о лучшей производительности прогнозирования. Очевидно, что производительность XGBoost является лучшей среди всех методов, использующих разные периоды времени, потому что она превосходит другие методы по 4 из 5 показателей, которые получают наименьшие MAPE и MSE и самые высокие оценки и оценки в трех группах экспериментов. В то время как для разных значений Δt существуют различные результаты прогнозирования. Производительность $\Delta t = 7$ достигает наилучшего результата, а результаты $\Delta t = 10$ являются худшими. Однако существует лишь небольшая разница между результатами $\Delta t = 5$ и $\Delta t = 7$. В следующих частях мы анализируем результаты $\Delta t = 7$ в наборе данных MAG.

Прогностическая эффективность этих методов для набора данных Американского физического общества APS показана в табл. 3. В таблице мы заметили, что общие показатели прогнозирования на точках доступа лучше, чем на MAG, что имеет меньшие погрешности и более высокую степень подгонки (R^2). Мы заметили, что производительность XGBoost также является лучшей. В отличие от результатов по MAG, все методы работают лучше на периоде $\Delta t = 10$, чем в других группах. В следующих частях мы анализируем результаты при $\Delta t = 10$ в наборе данных APS.

Таблица 2

Производительность алгоритмов разностного обучения на MAG

		MAE	MAPE	MSE	ACC	(R^2)
$\Delta t=5$	XGBoost	0,73	0,07	1,09	0,86	0,99
	LR	0,82	0,10	1,18	0,80	0,92
	GBDT	0,69	0,08	1,16	0,84	0,95
	CART	0,96	0,18	2,30	0,79	0,81
$\Delta t=7$	XGBoost	0,79	0,07	1,19	0,86	0,99
	LR	0,83	0,11	1,30	0,79	0,90
	GBDT	0,73	0,09	1,25	0,85	0,94
	CART	0,98	0,20	2,49	0,78	0,80
$\Delta t=10$	XGBoost	0,81	0,09	1,27	0,83	0,91
	LR	0,84	0,13	1,43	0,73	0,86
	GBDT	0,74	0,10	1,32	0,81	0,84
	CART	0,99	0,29	2,68	0,74	0,63

Производительность алгоритмов разностного обучения на точках доступа APS

		MAE	MAPE	MSE	ACC	(R ²)
$\Delta t=5$	XGBoost	0,54	0,06	0,62	0,81	0,97
	LR	0,57	0,08	0,65	0,80	0,96
	GBDT	0,55	0,06	0,63	0,80	0,96
	CART	0,56	0,09	0,73	0,78	0,95
$\Delta t=7$	XGBoost	0,51	0,07	0,55	0,82	0,97
	LR	0,53	0,10	0,55	0,80	0,96
	GBDT	0,51	0,07	0,56	0,81	0,97
	CART	0,54	0,08	0,70	0,80	0,96
$\Delta t=10$	XGBoost	0,45	0,06	0,46	0,85	0,98
	LR	0,47	0,07	0,48	0,83	0,97
	GBDT	0,45	0,08	0,47	0,84	0,97
	CART	0,46	0,07	0,60	0,80	0,96

Оценка причинно-следственных факторов

Приведенный выше анализ подтверждает причинно-следственные связи между различными факторами и общими результатами h -индекса. Однако вклад и важность факторов все еще нуждаются в изучении. Чтобы решить этот вопрос, мы сначала применили метод "складного ножа" [28] для проверки функции факторов каждой группы в отдельности. Метод "складного ножа" включает в себя две фазы: Добавление и удаление. На этапе добавления мы каждый раз используем одну группу факторов для прогнозирования результата. На этапе удаления мы удаляем группу факторов и обучаем модель с помощью остальных факторов. После этих двух этапов можно изучить индивидуальный вклад фактора в общую задачу прогнозирования.

Как показано на рис. 5, в экспериментах с набором данных MAG снижение значений ACC за счет удаления факторов, ориентированных на статьи, в четырех методах демонстрирует, что они имеют большое значение для прогнозирования h -индекса. Напротив, при устранении других типов факторов снижение прогностической эффективности не так очевидно. Этот факт может показать важность факторов, ориентированных на статьи, для прогнозирования будущего успеха ученых. Что касается дополнительных факторов, то факторы, ориентированные на статьи, по-прежнему играют важную роль в прогнозировании будущего научного влияния. Более того, факторы, ориентированные на автора и факторы времени, также демонстрируют свою эффективность для прогнозирования h -индекса ученого.

Те же анализы выполняются для набора данных APS, как показано на рис. 6. В отличие от результатов для MAG, снижение значений ACC из-за удаления факторов, ориентированных на автора, сильно влияет на прогнозирование h -индекса, что указывает на то, что факторы, ориентированные на

автора, имеют большое значение в наборе данных APS. Как и в предыдущих экспериментах, при устранении других типов факторов снижение прогностической эффективности не столь очевидно. Что касается факторов добавления, то факторы, ориентированные на автора, по-прежнему играют свою важную роль. Более того, факторы, ориентированные на статьи, также показывают свою эффективность для прогнозирования h -индекса ученого в наборе данных APS, но другие функции не имеют очевидных эффектов, что отличается от результатов эксперимента на MAG.

Кроме того, мы проанализировали важность функции, придаваемую XGBoost результатам в наборе данных MAG и APS. В XGBoost важность объекта может быть рассчитана как время, в течение которого объект использовался для разделения выборок на листьях деревьев в модели. Чем чаще использовалась функция, тем важнее это для модели. Как показано на рис. 7, в обоих наборах данных основные влияющие факторы по-прежнему совпадают с приведенными выше выводами, где факторы, ориентированные на статьи, по-прежнему являются наиболее важными для прогнозирования будущего академического успеха в MAG, который набирает 41,67% баллов важности; а факторы, ориентированные на автора, важны в APS, который набирает 42,33% баллов по важности.

Помимо этого, мы также анализируем коэффициент Джини различных учреждений. Коэффициент Джини, меньший 0,2, указывает на абсолютное равенство учреждений. Коэффициент Джини от 0,2 до 0,3 указывает на то, что учреждения относительно равны. Коэффициент Джини от 0,3 до 0,4 указывает на то, что учреждения относительно «нормальны». Наконец, коэффициент Джини, превышающий 0,4, указывает на большое неравенство между учреждениями.

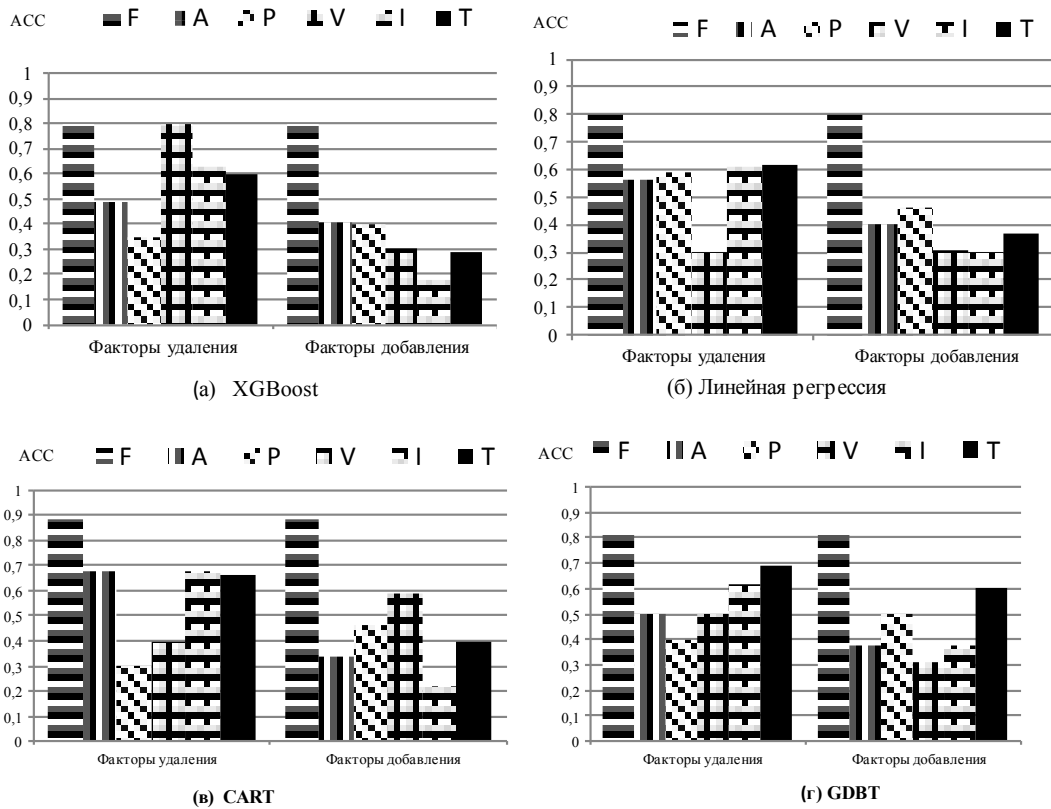


Рис. 5. Анализ факторного вклада на MAG. Четыре модели, обученные только с указанными факторами или без них. F: полный набор функций; A: Факторы автора; P: Факторы статьи; V: Факторы места проведения; I: Факторы учреждения; T: Факторы времени.

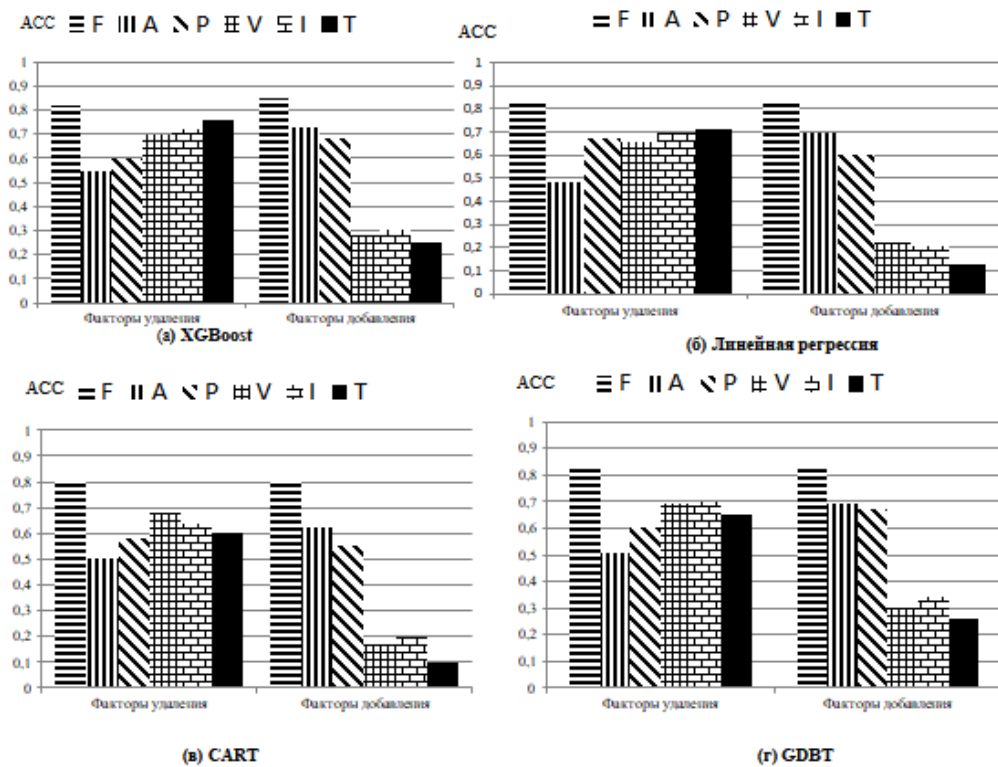


Рис. 6. Анализ факторного вклада на APS. Четыре модели, обученные только с указанными факторами или без них. F: полный набор функций; A: факторы автора; P: факторы статьи; V: факторы места проведения; I: факторы учреждения; T: факторы времени.

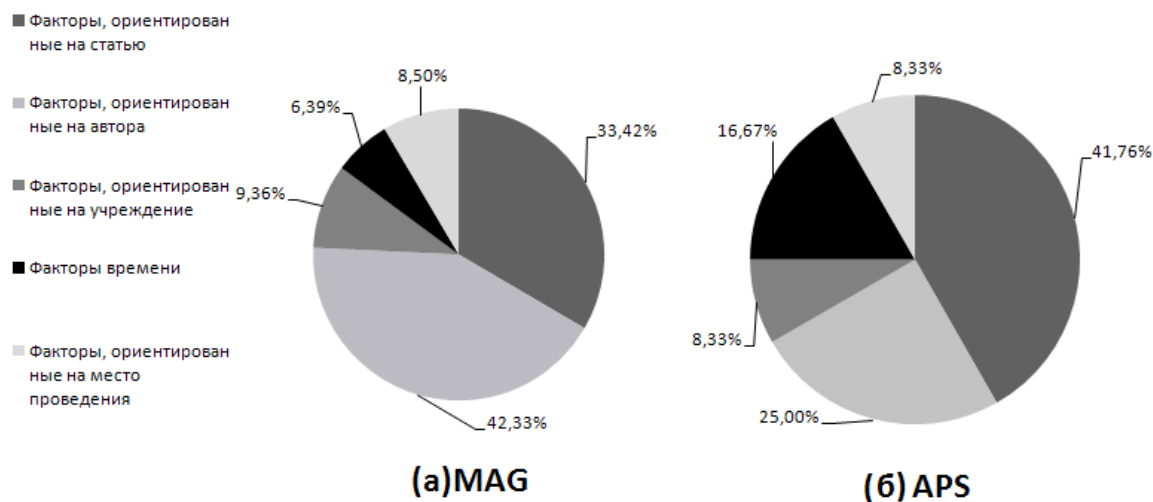


Рис. 7. Оценка важности различных факторов

Таблица 4

Средние коэффициенты Джини высших учебных заведений по MAG

	Количество статей	Цитирование	h-индекс
Наивысшие 5 %	0,102418	0,161101	0,042667
Наивысшие 10 %	0,191524	0,277041	0,091791
Наивысшие 20 %	0,230564	0,327122	0,121142
Наименьшие 10 %	0,351485	0,452952	0,200423

Таблица 5

Средние коэффициенты Джини высших рейтинговых учреждений Американского физического общества (APS)

	Количество статей	Цитирование	h-индекс
Наивысшие 5 %	0,011092	0,070037	0,015215
Наивысшие 10 %	0,169559	0,205179	0,145778
Наивысшие 20 %	0,216014	0,249943	0,184754
Наименьшие 10 %	0,266891	0,297307	0,228714

Чтобы получить всеобъемлющий коэффициент Джини для учреждения, мы сначала ранжируем учреждения в соответствии с количеством ученых. Затем в соответствии с рейтинговым списком можно получить коэффициент Джини по цитированию, количеству публикаций и h -индексу. Мы показываем коэффициент Джини для первых 5%, 10%, 20% и последних 10%. Как показано в таблицах 4 и 5, существуют некоторые интересные явления. Для топ-5% рейтинговых учреждений, как в MAG, так и в APS, их значения коэффициента Джини составляют менее 0,2, что указывает на то, что h -индекс ученых очень близок к их коллегам в том же учреждении. В топ-10% рейтинговых институтов, за исключением цитирования, коэффициенты

Джини для количества статей и h -индекса составляют менее 0,2, что по-прежнему показывает равенство ученых по этим двум аспектам. В то время как для учреждений, входящих в топ-20% и последние 10%, их коэффициент Джини для количества статей и цитирования превышает 0,2. Очевидно, что существуют некоторые различия в количестве статей и цитируемости исследователей в этих учреждениях. Однако уровень h -индекса ученых во всех упомянутых выше учреждениях очень похож на уровень их коллег. Это явление показывает, что ученые в одном и том же учреждении – это люди одного склада. И это явление одинаково в информатике и физике. Причина этого заключается в том, что научное общение между ними очень

удобное и частое, и они могут в какой-то степени непосредственно ощущать давление со стороны своих коллег. Поэтому ученые стараются не отстаивать от своих коллег в академических исследованиях, и их общее научное воздействие весьма похоже друг на друга. Кроме того, при предоставлении преподавательских должностей для исследователей могут существовать стандартные требования к найму для одного и того же учреждения. Как следствие, ученые в одном и том же учебном заведении находятся на одном и том же академическом уровне.

ЗАКЛЮЧЕНИЕ И ДАЛЬНЕЙШАЯ РАБОТА

В этой статье мы стремимся выявить причинно-следственные факторы, которые играют решающую роль в прогнозировании академических успехов ученых. Чтобы решить эту проблему, мы сначала предлагаем пять потенциальных причинно-следственных факторов, которые являются факторами, ориентированными на статью, факторами, ориентированными на автора, факторами, ориентированными на место проведения, факторами, ориентированными на учреждение, и факторами времени. Затем, используя самые современные алгоритмы машинного обучения, мы обнаружили, что факторы, ориентированные на автора и статью, являются наиболее значимыми причинно-следственными факторами для прогнозирования будущего успеха ученых.

Кроме того, мы анализировали вклад каждого фактора, используя метод "складного ножа" и оценивая факторы в процессе прогнозирования. Результаты еще раз демонстрируют важность факторов, ориентированных на статью и автора. Далее мы анализировали конкретный рейтинг важности этих пяти групп факторов, используемых в наших экспериментах. После этого процесса мы обнаружили, что в наборе данных MAG факторы, ориентированные на статью, имеют важность 41,47%, факторы, ориентированные на автора, имеют важность 25%, факторы, ориентированные на время, имеют важность 16,67%, а факторы, ориентированные на место и учреждение, имеют важность 8,33%, в то время как в наборе данных APS факторы, ориентированные на статью, имеют важность 33,42%, факторы, ориентированные на автора, имеют важность 42,33%, факторы, ориентированные на время, имеют важность 6,39%, факторы, ориентированные на место, имеют важность 8,50%, а факторы, ориентированные на учреждение, имеют важность 9,36%. Между тем, мы также обнаружили, что *h*-индекс ученых в одних и тех же учреждениях, как правило, очень близок друг к другу.

В дальнейшем мы планируем выявить больше факторов и провести наши эксперименты с другими наборами данных из различных дисциплин, чтобы продемонстрировать обоснованность нашей работы.

ЛИТЕРАТУРА

1. *Acuna D. E., Allesina S., Kording K. P.* Future impact: Predicting scientific success // *Nature*. — 2012. — Vol. 489, No. 7415. — P. 201.
2. *Amjad T., Daud A., Che D., AtiaAkram.* MuICE: Mutual influence and citation exclusivity author rank // *Information Processing & Management*. — 2016. — Vol. 52, No 3. — P. 374–386.
3. *Amjad T., Ding Y., Daud A., Xu J., Malic V.* Topic-based heterogeneous rank // *Scientometrics*. — 2015. — Vol. 104, No. 1. — P. 313–334.
4. *Amjad T., Ding Y., Xu J., Zhang C., Daud A., Tang J., Min Song M.* Standing on the shoulders of giants // *Journal of Informetrics*. — 2017. — Vol. 11, No. 1. — P. 307–323.
5. *Bai X., Ivan Lee I., Ning Z., Tolba A., Xia F.* The role of positive and negative citations in scientific evaluation // *IEEE Access*. — 2017. — Vol. 5, No. 99. — P. 17607–17617.
6. *Bornmann L., Haunschild R.* How to normalize Twitter counts? A first attempt based on journals in the Twitter index // *Scientometrics*. — 2016. — Vol. 107, No. 3. — P. 1405–1422.
7. *Cao X., Yan Chen Y., Ray Liu K. J.* A data analytic approach to quantifying scientific impact // *Journal of Informetrics*. — 2016. — Vol. 10, No. 2. — P. 471–484.
8. *Clauset A., Larremore D. B., Sinatra R.* Data-driven predictions in the science of science // *Science*. — 2017. — Vol. 355, No. 6324. — P. 477–480.
9. *Costas R., Zahedi Z., Wouters P.* Do "altmetrics" correlate with citations? Extensive comparison of altmetric indicators with citations from a multidisciplinary perspective // *Journal of the Association for Information Science & Technology*. — 2015. — Vol. 66, No 10. — P. 2003–2019.
10. *Dong Y., Johnson R., Chawla N.* Can scientific impact be predicted? // *IEEE Transactions on BigData*. — 2016. — Vol. 2, No. 1. — P. 18–30.
11. *Dunaiski M., Visser W., Geldenhuys J.* Evaluating paper and author ranking algorithms using impact and contribution awards // *Journal of Informetrics*. — 2016. — Vol. 10, No. 2. — P. 392–407.
12. *Egghe E.* Theory and practise of the g-index // *Scientometrics*. — 2006. — Vol. 69, No. 1. — P. 131–152.
13. *Fortunato S., Bergstrom C. T., Börner K., J. A. Evans J. A., Helbing D., Milojević S., Petersen A. M., Radicchi F., Sinatra R., Uzzi B.* Science of science // *Science*. — 2018. — Vol. 359, No. 637. — P. 1007–1007.
14. *Garfield E.* The history and meaning of the journal impact factor // *Jama*. — 2006. — Vol. 295, No. 1. — P. 90–93.
15. *Heiberger R. H., Wiecek O. J.* Choosing collaboration partners. How scientific success in Physics depends on network positions. — 2016. — [arXiv:1608.03251 abs/1608.03251 (2016), p.201–207].
16. *Hirsch J. E.* An index to quantify an individual's scientific research output // *Proceedings of the Nation-*

al academy of Sciences of the United States of America. — 2005. — Vol. 102, No. 46. — P. 16569–16572.

17. Kleinberg J. M. Authoritative sources in a hyper-linked environment // J. ACM. — 1999. — Vol. 46, No. 5 (Sept. 1999). — P. 604–632.

18. Klimek P., Jovanovic A. S., Egloff R., Schneider R. Successful fish go with the flow: Citation impact prediction based on centrality measures for term document networks // Scientometrics. — 2016. — Vol. 107, No. 3. — P. 1265–1282.

19. Kong X., Shi Y., W., Kai Ma, Wan L., Xia F. The evolution of Turing Award Collaboration Network: Bibliometric-level and network-level metrics // IEEE Transactions on Computational Social Systems. — 2019. — Vol. 6, No. 6. — P. 1318–1328. — <https://doi.org/10.1109/TCSS.2019.2950445>

20. Le T., Liu L., Tsykin A., Goodall G. J., Liu B., Sun B. -Y., Li J. Inferring microRNA-mRNA causal regulatory relationships from expression data // Bioinformatics. — 2013. — Vol. 29, No. 6. — P. 765–771.

21. Li J., Le T., Liu L., Liu J., Jin Z., Sun B., Ma S. From observational studies to causal rule mining // ACM Transactions on Intelligent Systems and Technology (TIST). — 2016. — Vol. 7, No. 2. — P. 14.

22. Li L., Tong H. The child is father of the man: Foresee the Success at the early stage / AcmSigkdd International Conference on Knowledge Discovery & Data Mining// ACM, Sydney, NSW. — 2015. — P. 655–664.

23. Liang R., Jiang X. Scientific ranking over heterogeneous academic hyper-network // Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence. AAAI, AAAI Press, Phoenix, Arizona, USA. — 2016. — P. 20–26.

24. Liu J., Xia F., Wang L., Xu B., Kong X., Tong H., King I. Shifu2: A network representation learning based model for advisor-advisee relationship mining // IEEE Transactions on Knowledge and Data Engineering. — 2019. — Vol. 33, No. 4. — P. 1763–1777. — <https://doi.org/10.1109/TKDE.2019.2946825>

25. Liu L., Wang Y., Sinatra R., Lee Giles C., Song C., Wang D. Hot streaks in artistic, cultural, and scientific careers // Nature. — 2018. — Vol. 559, No. 7714. — P. 396–399. — <https://doi.org/10.1038/s41586-018-0315-8>

26. Liu Y., Zhang L., Nie L., Yan Y., Rosenblum D. S. Fortune teller: Predicting your career path / Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence. — 2016. — Vol. 2016. — AAAI Press, Phoenix, Arizona, USA. — P. 201–207.

27. Page L., Brin S., Motwani R., Winograd T. The PageRank citation ranking: Bringing order to the Web. Technical Report 1999-66. — Stanford InfoLab, 1999. — [http://ilpubs.stanford.edu:8090/422/Previous number =SIDL-WP-1999-0120](http://ilpubs.stanford.edu:8090/422/Previous%20number%20SIDL-WP-1999-0120).

28. Severiano A., Carriço J. A., Robinson D. A., Ramirez M., Pinto F. R. Evaluation of jackknife and bootstrap for defining confidence intervals for pairwise agreement measures // PLoS One. — 2011. — Vol. 6, No. 5, e19539.

29. Shiffrin R. M. Drawing causal inference from Big Data // Proceedings of the National Academy of Sciences. — 2016. — Vol. 113, No. 27. — P. 7308–7309.

30. Sinatra R., Wang D., Deville P., Song C., Barabási A.-L. Quantifying the evolution of individual scientific impact // Science. — 2016. — Vol. 354, No. 6312. — P. aaf5239–aaf5239.

31. Spirtes P., Zhang K. Causal discovery and inference: Concepts and recent methodological advances // Applied Informatics. — 2016. — Vol. 3, No. 1 (02 2016). — P. 1–28.

32. Stegehuis C., Litvak N., Waltman L. Predicting the long-term citation impact of recent publications // Journal of Informetrics. — 2015. — Vol. 9, No.3. — P. 642–657.

33. Valenzuela M., Ha V., Etzioni O. Identifying meaningful citations / Workshops at the Twenty-Ninth AAAI Conference on Artificial Intelligence, Vol. WS-15-13. — AAAI, AAAI Press, Austin, Texas, USA, 2015. — P. 21–26.

34. Van Houten B., Phelps J., Barnes M., Suk W. A. Evaluating scientific impact // Environmental health perspectives. — 2000. — Vol. 108, No. 9. — P. A392–A393.

35. Wang D., Chaoming Song C., Barabási A.-L. Quantifying long-term scientific impact // Science. — 2013. — Vol. 342, No. 6154. — P. 127–132.

36. Wang S., Xie S., Zhang S., Li Z., He Y., He Y. Coranking the future influence of multiojects in bibliographic network through mutual reinforcement // Acm Transactions on Intelligent Systems & Technology. — 2016. — Vol. 7, No. 4. — P. 64:1–64:28.

37. Wang W., Yu S., Bekele T. M., Kong X., Xia F. Scientific collaboration patterns vary with scholars' academic ages // Scientometrics. — 2017. — Vol. 112, No. 1. — P. 329–343.

38. Wu H., Sun M., Mi P., Tatti N., Chris North C., Ramakrishnan N. Interactive discovery of coordinated relationship chains with maximum entropy models // ACM Transactions on Knowledge Discovery from Data (TKDD). — 2018. — Vol. 12, No. 1. — P. 7.

39. Xia F., Su X., Wei W., Zhang C., Ning Z., Lee I. Bibliographic analysis of Nature based on Twitter and Facebook altmetrics data // Plos One. — 2016. — Vol. 11, No. 12.

40. Xia F., Wang W., Bekele T. M., Liu H. Big Scholarly Data: A survey// IEEE Transactions on Big Data. — 2017. — Vol. 3, No. 1. — P. 18–35.

41. Yang Y., Tang J., Li J. Learning to infer competitive relationships in heterogeneous networks // ACM Transactions on Knowledge Discovery from Data (TKDD). — 2018. — Vol. 12, No.1. —P. 12.

42. Yu D., Wang W., Zhang S., Zhang W., Liu R. A multiple-link, mutually reinforced journal-ranking model to measure the prestige of journals // Scientometrics. — 2017. — Vol. 111, No.1. — P. 521–542.

43. Zhang J., Ning Z., Bai X., Kong X., Zhou J., Xia F. Exploring time factors in measuring the scientific impact of scholars // Scientometrics. — 2017. — Vol. 112, No. 3. — P. 1301–1321.

Мы стоим на ненадежных плечах? Влияние цитирования отозванных из публикации статей на предыдущие и последующие опубликованные статьи: исследование базы данных Web of Science*

Сепиде ФАХИМИФАР
(Sepideh FAHIMIFAR)

Отделение информации и науки о знаниях,
факультет менеджмента Тегеранского
университета, г. Тегеран, Иран

Али ГОРБИ
(Ali GHORBI)

Факультет менеджмента Тегеранского
университета, г. Тегеран, Иран

Марсель АУСЛООС
(Marcel AUSLOOS)

Школа бизнеса Лестерского университета,
Брукфилд, г. Лестер, Великобритания

Кафедра статистики и эконометрики,
Бухарестский университет экономических
исследований, КалеаДоробантилор,
г. Бухарест, Румыния

В настоящем исследовании предпринята попытка определить влияние отозванных из публикации статей на предыдущие или последующие статьи. Мы рассматриваем 5693 отозванных документа с 1975 по 2020 год индексированных в базе данных Web of Science на основе библиометрических методов. В качестве технических средств мы используем программное обеспечение HistCite, Excel и SPSS. Полученные результаты свидетельствуют о существенной разнице между средним количеством отозванных и поданных на публикацию статей при цитировании в отозванных статьях. Кроме того, существует значительная разница между средним количеством поданных и отозванных статей, цитирующих отозванные статьи. Причинами отзыва статьи могут быть не предыдущие отозванные статьи, однако поданные на публикацию статьи могут быть отозваны позже из-за ссылок на (многие) отозванные статьи. Делается вывод, что владельцы баз данных цитирования должны тщательно сосредоточиться на этих статьях, проверяя ссылки на каждую новую статью, в которой цитируются ранее отозванные статьи.

ВВЕДЕНИЕ

Публикации в научных журналах продолжают маркировать состояние прогресса в научном сообществе [1]. Однако публикация научной статьи иногда отклоняется от ее первоначального пути. Это заканчивается отзывом из публикации, самой серьезной стигматизацией в научных исследованиях [2], помимо обязательных ссылок, приводя к аномальным показателям качества исследований [3] и плагиату [4]. Отзыв из публикации — это публичное уведомление о том, что статья должна быть отозвана из-за ошибок или необоснованных данных. Довольно тревожно, что в последнее время значительно возросло количество отозванных доку-

ментов [1]. Однако публикация научной статьи иногда отклоняется от ее первоначального пути. Это заканчивается отзывом из публикации, самой серьезной стигматизацией в научных исследованиях [2], помимо обязательных ссылок, приводя к аномальным показателям качества исследований [3] и плагиату [4]. Отзыв из публикации — это публичное уведомление о том, что статья должна быть отозвана из-за ошибок или необоснованных данных. Довольно тревожно, что в последнее время значительно возросло количество отозванных доку-

* Перевод Fahimifar S., Ghorbi A., Ausloos M. Are we standing on unreliable shoulders? The effect of retracted papers citations on previous and subsequent published papers: A study of the Web of Science database// International Journal of Information Science and Management. — 2022. — Vol. 20, No. 1. — P. 319-333. — <https://arxiv.org/ftp/arxiv/papers/2201/2201.09090.pdf>

ментов [5]. Изучая статьи на тему отзыва из публикации, мы можем констатировать, что исследования по отозванным статьям в основном рассматривали:

1. Причины отзыва: из-за ошибки издателя, ошибки авторов, ранее не учтенной ошибки, самоцитирования, почетного авторства (с указанием имен тех, кто не играл роли или, по крайней мере, играл второстепенную роль) [6-12].

2. Увеличение доли отозванных статей [13-16] до такой степени, что причина такого поведения кажется неясной [3; 17, 18]; некоторые причины связаны с экономическими аспектами, присущими нынешней системе развития науки, такими как гранты, премии, финансирование, рабочие места и т.д. [19, 20]. С другой стороны, быстрая скорость обнаружения отозванных статей коренится в самовыражении авторов [21], создании некоторых веб-сайтов для отслеживания исследований, таких как *retraction watch** [10], повышении уровня контроля над статьями благодаря широкому использованию Интернета, а также углубленному обзору некоторыми главными редакторами журналов или рецензентами [11].

3. Влияние отозванных статей на будущие научные результаты путем изучения различных видов цитирований, в которых большинство ссылок были отмечены как положительные (без какой-либо критики) [22-25].

4. Также учитывался временной интервал между публикацией статьи и ее отзывом [2, 7, 18]. В среднем требуется около двух лет, чтобы статья была отозвана после публикации [26-29; 25].

5. Влияние цитирования на отозванные статьи было одной из наиболее обсуждаемых тем. Имеющиеся данные свидетельствуют о том, что ссылки на отозванные статьи продолжают, как если бы они были подлинными работами [30, 19]. Интересно, что цитирование этих статей может иметь положительный, отрицательный или нейтральный эффект [26] до или после отзыва статьи [31, 32]. Однако политика редакции и издателя может отличаться и меняться со временем [33]. Положительные эффекты учитываются, когда цитаты используют содержание отозванных статей для поддержки аргументации, независимо от научной обоснованности этих отозванных статей. Негативные последствия, напротив, возникают, когда цитаты критикуют содержание отозванных статей (до или после отзыва) и подчеркивают ошибки, допущенные в этих статьях. Конечно, нейтральные эффекты относятся к цитатам, которые ссылаются на отозванную (или подлежащую отзыву из публикации) статью в разделе литературы без какой-либо оценки относительно ее достоверности. Бар-Илан и Халеви [26] показали, что 83% ссылок на отозванные статьи были положительными, 12% - нейтральными и только 5% - отрицательными.

Бар-Илан и Халеви [26] также предположили, что положительное цитирование отозванных статей продолжает появляться, несмотря на уведомление об отзыве на онлайн-платформах издателя и без внимания на причины отзыва. Более того, учитывая, что обычно требуется два года, чтобы статья была отозвана после ее публикации [26-29, 25], есть хорошие шансы, что отозванная статья будет влиять на ее последующие статьи. Кроме того, многие отозванные статьи, вероятно, продолжают получать ссылки после отзыва [2]. Например, многие отозванные статьи в области радиационной онкологии, по-видимому, цитируются после их отзыва; на многие ссылались как на подлинные статьи [34]. Другое исследование отозванных и поданных статей в области инженерии показывает, что их цитирование продолжается и влияет на их научную кредитоспособность [35]. Кроме того, некоторые риски возникают после продолжения цитирования отозванных документов. Некоторые авторы считают, что их выводы и результаты все еще обоснованы [36], и это, в некотором смысле, может ввести читателей в заблуждение.

Это означает, что такого рода статьи, написанные с использованием фальсификации, подделки или недостоверных данных, могут привести к сомнительным результатам. Для этих результатов продвижение несколько обманчиво, когда мы используем их в качестве ориентира или при решении проблем. Таким образом, количественное знание отозванных статей, цитируемых в последующих публикациях, может послужить сигналом тревоги для нашего научного сообщества и широкой общественности. Помимо изучения вероятности отзыва, еще одна цель нашей статьи - найти (и отобразить) причину (причины) и влияние (последствия) отозванной статьи на последующие публикации. Мы считаем, что, публикуя такие предположения, мы стремимся избежать включения библиографической информации об отозванных статьях в поданные статьи; тем самым, мы надеемся, что сможем помочь ускорить рост истинной науки. А значит наше исследование направлено на измерение влияния отозванных статей на научные публикации. Наш анализ основан на случаях из журналов, индексируемых в базе данных WoS. Это самая важная база данных во всем мире, охватывающая различные дисциплины и позволяющая проводить сравнения в различных научных областях.

Во-первых, мы отслеживаем протяженность во времени числа отозванных документов. Учитывая множество вновь появляющихся журналов, следуя схеме открытого доступа, мы добавляем в число предположений "Ответственность об исследованиях" измерения возможности существования значительной корреляции между квантилями журнала "а" и количеством отозванных статей. Измерение корреляции между качеством журналов и количеством отозванных статей связано с тем фактом, что

* <http://retractiondatabase.org/RetractionSearch.aspx?>

последствия для исследовательской деятельности могут быть более серьезными, если отозванные статьи будут из ведущих журналов, поскольку большинство исследователей и аудиторий могут полагаться на их выводы. Действительно, нет необходимости подчеркивать, что любое текущее исследование очень сильно зависит от предыдущих исследований. В результате более ранние статьи играют заметную роль в формировании тех, которые будут опубликованы позже. Таким образом, интересно изучить, как часто цитируются отозванные статьи и какие из них цитируются отозванными работами, как они влияют на будущие исследования!

Таким образом, настоящее исследование направлено на то, чтобы ответить на следующие вопросы:

- Сколько статей было отозвано (с 1975 по 2020 год)?

- Каково распределение отозванных статей в журналах, индексируемых в базе данных Journal Citation Report database (JCR)?

- Каково влияние отозванных статей на статьи журналов JCR (отозванные или нет)?

Мы также хотим достичь некоторых исследовательских целей, включая следующие:

1. Исследовать, есть ли существенная корреляция между квартилями журналов (Q1, Q2, Q3 и Q4) и количеством отозванных статей, т.е. были ли отозванные статьи в значительной степени опубликованы в высококачественных журналах с высокими импакт-факторами.

2. Выяснить, есть ли существенная разница между средним числом поданных и отозванных статей, цитирующих отозванные статьи; на самом деле, мы хотим знать, цитируются ли отозванные статьи чаще, чем отозванные статьи, по сравнению с поданными статьями.

3. Выяснить, отмечается ли существенная разница между средним количеством отозванных и поданных статей, цитируемых отозванными статьями; т.е. отозванные статьи являются результатом предыдущих поданных статей или других отозванных статей.

МЕТОДОЛОГИЯ

Мы собрали данные из базы данных WoS (7 июня 2020 года*) в соответствии со стратегией поиска, описанной на рис. 1. Данные включали общее количество документов и цитируемых ссылок в период с 1975 по 2020 год. Уведомления об отзыве были исключены во избежание дублирования.

Мы должны заметить, что в WoS не все отозванные статьи проиндексированы по типу публикации "Отозванная публикация". Согласно WoS, отозванная публикация – это статья, которую отозвал автор, учреждение, редактор или издатель из-за ошибок или необоснованных данных. Однако мы используем данные из WoS вместо базы данных RetractionWatch (retractiondatabase.org) из-за нашего выбора программного обеспечения. При этом, собирая данные из индексов цитирования, таких как WoS и Scopus, мы можем импортировать данные в программное обеспечение по наукометрии и библиометрии.

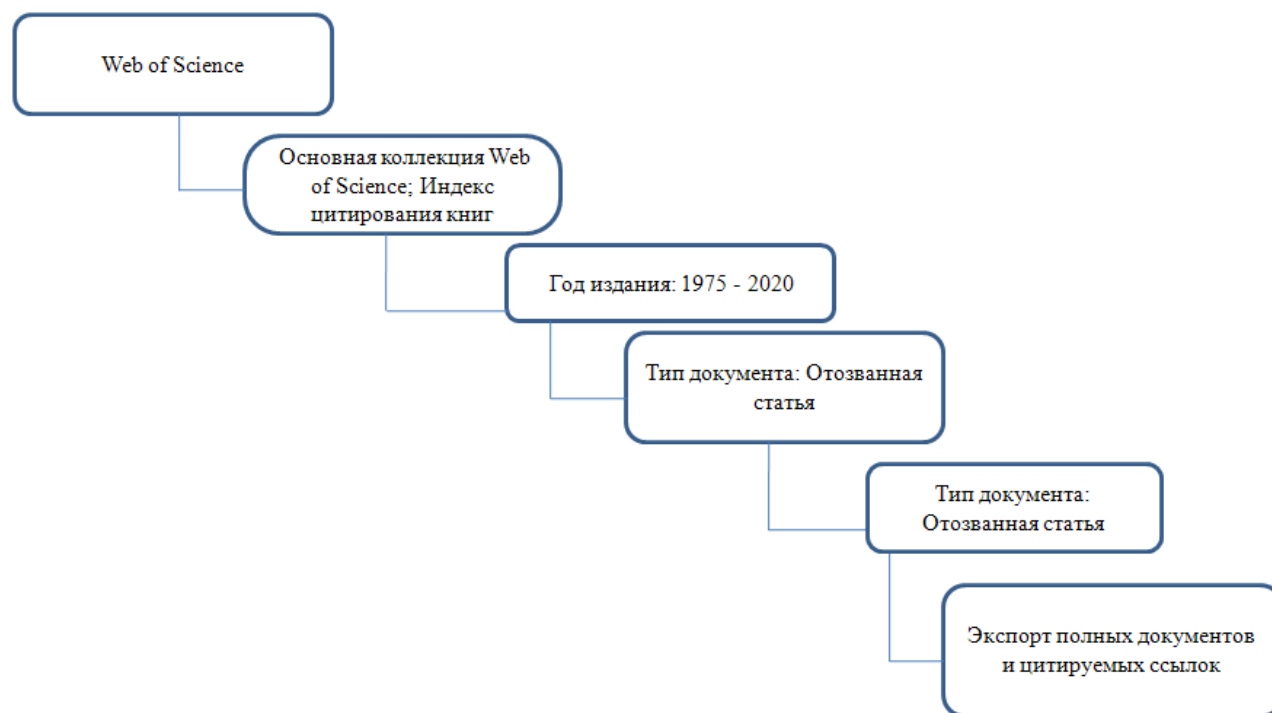


Рис. 1. Стратегия поиска для сбора данных

* В базу данных WOS был отправлен запрос, позволяющий нам получить все публикации, опубликованные в период с 1975 по 2020 год без каких-либо ограничений в зависимости от типа документа.

Мы не используем RetractionWatch, потому что большинство программ по библиометрии принимают только выходные данные из некоторых совместимых баз данных цитирования. В этом исследовании мы также игнорируем документы с тегом "Ретракция". К ним относятся некоторые документы, которые были помечены как отзыв из публикации в WoS, но статья не была отозвана, т.е. практически не связана с журналом или веб-сайтом. Примером может служить одна статья, озаглавленная "Основанный на доказательствах подход к индикаторам для проведения предварительных оценок воздействия развития гидроэнергетики на окружающую среду". Некоторые редакторы преждевременно опубликовали упомянутую статью, но правильная версия этой статьи была опубликована в другом журнале. Более того, есть некоторые документы с тегами "отзыв" без каких-либо следов полного текста статьи, даже с подписями "отозванная статья". Обратите внимание, что некоторые уведомления об отзыве можно найти в WoS; однако полный текст соответствующих статей неожиданно не помечен как "отозванная статья". Примером может служить "микроРНК-493 подавляет рост опухоли, инвазию и метастазирование рака легкого, регулируя E2F1"

Мы нашли 5693 отозванных статьи 19482 авторов, опубликованных в 2046 журналах. После экспорта данных из WoS в виде текстового файла мы обеспечили данным проверку первой строки текстового файла и заменили ее фразой "Экспорт формата ISI". Затем мы импортировали собранные данные из WoS в программное обеспечение HistCite. После этого мы экспортировали выходные данные из программного обеспечения HistCite™ в формат Excel в виде файла ".csv". Однако, используя заголовки фильтрации и наблюдая за содержимым каждой ячейки, мы поняли, что некоторые данные были неправильно отсортированы при передаче. В результате мы проверили документы один за другим, чтобы убедиться, что данные указаны правильно, в соответствии с

- (i) количеством авторов каждой отозванной статьи;
- (ii) количеством ссылок на каждую статью во всей базе данных WoS (GCS)**;
- (iii) количеством ссылок на каждую отозванную статью в других отозванных статьях (LCS)***;
- (iv) количеством поданных и отозванных статей в разделе ссылок каждой отозванной статьи.

*Чтобы не увеличивать видимость подобного рода статей, читатель может отметить, что мы не явно ссылаемся на них в списке литературы для чтения.

** GCS- Global Citation Score

*** LCS- Local Citation Score

Извлеченные данные затем вводились в программное обеспечение SPSS (версия 21.0; SPSS) для статистического анализа.

Затем, чтобы ответить на исследовательский вопрос о качестве журналов, в которых были отозваны статьи, был извлечен список журналов вместе с квартилями, опубликованными базой данных JCR. После этого мы удалили дублированные заголовки. Действительно, журнал может быть отнесен к разным квартилям из-за различных областей исследований, которые он охватывает. Мы экспортировали названия журналов, связанных с отозванными статьями, в виде CSV-файла. Наконец, мы сравнили названия журналов и списки журналов с квартилями, чтобы присвоить каждому журналу уникальный квартиль таким образом, чтобы журнал с наивысшим квартилем был сохранен.

РЕЗУЛЬТАТЫ

Количество отозванных документов в год

На рис. 2 показано, сколько статей было отозвано с 1975 по 2020 год. За этот период увеличилось количество отозванных документов. Во-первых, оно оставалось относительно стабильным с 1975 по 1990 год. Однако после колебаний в период с 1991 по 1996 год можно наблюдать резкое увеличение с 1997 по 2007 год. В то время как количество отозванных документов снова колебалось в период с 2008 по 2013 год, оно достигло максимума в 453 отозванных документа. После этого их число уменьшилось, но не намного в последнее время: оно было все еще выше 50 в 2019 г. В XXI в. число отозванных статей растет, темп ровно удвоился с 2003 по 2010 гг., но с 2015 г. это число выравнивалось (рис. 2).

Распределение отозванных статей в журналах, индексируемых в базе данных JCR

Среди 2046 журналов*, в которых есть хотя бы одна отозванная статья, 1893 журнала имеют импакт-фактор и индексируются в JCR; только 153 журнала с 495 отозванными статьями не имеют импакт-фактора. В двух верхних квартилях насчитывается 1278 журналов, которые включают 3975 отозванных статей. Результаты табл. 1 свидетельствуют о том, что большинство отозванных статей были опубликованы в журналах с высоким импакт-фактором.

Это приводит к следующему разделу, в котором мы исследуем, существует ли корреляция между качеством журналов, измеряемым их квартильным показателем, и количеством отозванных статей в них.

* Помните, что мы ищем, появляется ли журнал в разных квартилях из-за его охвата научной областью, и мы выбираем лучший квартиль для журнала.

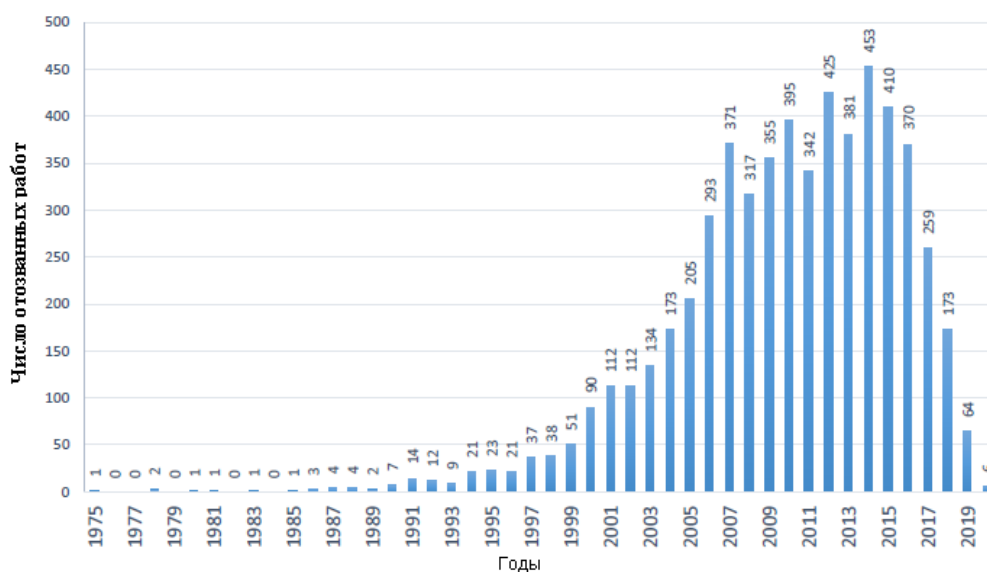


Рис. 2. Количество отозванных статей в год в журналах, индексируемых в WoS

Таблица 1

Распространение отозванных статей на основе рейтинга журнала в квартилях JCR

Общее количество отозванных документов: 5693								
Общее количество отозванных статей в журналах с импакт-фактором: 5198								
Квартили журналов	Q1		Q2		Q3		Q4	
Частота и происхождение	n	%	n	%	n	%	n	%
Отозванные статьи в журналах с импакт-фактором	2406	46%	1569	30%	753	15%	470	9%
Общее количество журналов с отозванными статьями: 2046								
Общее количество журналов с импакт-фактором, в которых были отозваны статьи: 1893								
Квартили журналов	Q1		Q2		Q3		Q4	
Частота и происхождение	n	%	n	%	n	%	n	%
Журналы с импакт-фактором, в которых отозваны статьи	736	39%	542	29%	386	20%	229	12%
Квартили журналов	Q1		Q2		Q3		Q4	
Процент отозванных статей в журнале с импакт-фактором (измеряется для количества отозванных статей из журнала в данном квартиле)	31%		35%		51%		48%	

Таблица 2

Корреляция между квартилями журналов и отозванными статьями

Переменная	Число	Среднее значение	Стандартное отклонение	Количество статей в каждом журнале	Квартильный показатель каждого журнала
1. Количество статей в каждом журнале	1,893	2,75	5,67	—	0,136*
2. Квартильный показатель каждого журнала	1,893	2,94	1,036	0,136*	—

* $p < .01$.

Корреляция между квартилями журналов и количеством отозванных статей

Чтобы определить, есть ли существенная корреляция между квартилем журнала и количеством отозванных статей, мы сначала рассчитываем количество отозванных статей в каждом журнале. Как упоминалось выше, если журнал появлялся более чем в одной дисциплине, использовался только самый высокий квартиль журнала. Корреляционный тест Спирмена используется для измерения переменных как в порядковой, так и непрерывной шкале. Коэффициент положительной корреляции (см. табл. 2) указывает на значительную положительную линейную корреляцию между рейтингом каждого журнала и количеством отозванных статей, хотя эта корреляция кажется очень слабой*. Таким образом, с уровнем достоверности 99%, можно сделать вывод, что первая гипотеза принята ($r_s [1893] = 0,136, p < 0,01$) (табл. 2).

Корреляция между средним количеством поданных и отозванных статей, цитирующих отозванные статьи

Давайте рассмотрим, относятся ли отозванные статьи в основном к другим ранее отозванным статьям или к поданным статьям. Другими словами, имеется ли какая-либо существенная разница между средним числом поданных и отозванных статей, в которых цитируются отозванные статьи.

Для этого мы используем U-критерий Манна-Уитни, как обычно делается при желании сравнить средние значения между двумя независимыми группами, когда зависимая переменная является либо порядковой, либо непрерывной, но не распределенной нормально [37]. Для этого рассчиты-

вается количество цитирований каждой отозванной статьи в базе данных WoS. Это приводит к общему количеству ссылок на статью (GCS) и общему количеству ссылок на отозванную статью другими отозванными статьями (LCS). Мы вычитаем количество ссылок на отозванные статьи (LCS) из общего числа ссылок (GCS) с учетом подсчета числа ссылок на отозванную статью относительно поданных статей. Во-первых, мы проверяем нормализацию данных. Результат теста Колмогорова-Смирнова показывает, что данные (со ссылкой на отозванные и поданные статьи) распределяются нормально, поскольку значение P меньше 0,01 (см. табл. 3). Поэтому следует использовать U-тест Манна-Уитни вместо независимого t-теста.

U-критерий Манна-Уитни (см. табл. 4) показывает, что количество ссылок, полученных отозванными статьями из поданных статей, больше, чем у других отозванных статей (см. средний ранг). Таким образом, имеются убедительные доказательства, подтверждающие разницу между цитатами, полученными в отозванных статьях, как из не отозванных (поданных), так и из других отозванных статей. Учитывая тот факт, что количество статей, в которых цитируются отозванные статьи, больше, чем количество статей, в которых цитируются поданные статьи, можно утверждать, что есть шанс, что эти цитируемые статьи будут считаться отозванными в будущем. Следовательно, если теоретические и содержательные дискуссии в рамках поданных статей исходят из содержания отозванных статей, эти обсуждения можно считать потенциально ненадежными. Тем не менее, для удовлетворительного вывода следует проанализировать содержание текста, чтобы определить, является ли цитата критикой или положительной оценкой.

Таблица 3

Тесты распределения нормальности для отозванных и поданных статей

Переменная	df	P-значение
1. Процитированные отозванные статьи	5693	0
2. Процитированные поданные статьи	5693	0

Таблица 4

Значительная средняя разница между цитированием отозванных и поданных статей

Переменная	Число	Средний ранг	P-значение
1. Процитированные отозванные статьи	5693	3046,86	0
2. Процитированные поданные статьи	5693	8340,14	0

* Корреляция — это мера того, как две или более переменных связаны друг с другом. Это диапазон от 0 до 1, включающий 00-.19 — очень слабый, 20 -.39 — слабый, 40 -.59 — умеренный, .60 -.79 — сильный и .80 -1,0 — очень сильный!.

Проверка нормальности числа ссылок для каждой отозванной статьи

Переменная	<i>df</i>	<i>P-значение</i>
1. Цитируемая отозванная статья	5693	0
2. Процитированные поданные статьи	5693	0

Таблица 6

Значительная средняя разница между количеством цитируемых отозванных и поданных статей

Переменная	<i>n</i>	<i>Средний ранг</i>	<i>P-значение</i>
1. Цитируемая отозванная статья	5693	2894,95	0
2. Процитированные поданные статьи	5693	8492,05	0

Разница между средним количеством отозванных и поданных статей, на которые ссылаются отозванные статьи

В этом разделе рассматривается, ссылаются ли авторы отозванных статей в основном на другие ранее отозванные статьи или в основном упоминают поданные статьи в своих публикациях. Есть ли какая-либо существенная разница между средним числом поданных и отозванных статей, на которые ссылаются отозванные статьи (т.е. отозванные статьи ссылаются на другие статьи), снова рассчитано с помощью U-критерия Манна-Уитни.

Следовательно, мы вычитаем количество отозванных ссылок из общего количества ссылок для подсчета числа поданных ссылок. Прежде всего, из-за данных о масштабе мы рассчитали нормализацию с помощью критерия Колмогорова-Смирнова (см. табл. 5).

U-критерий Манна-Уитни указывает на то, что большее количество ссылок, цитируемых в отозванных статьях, являются поданными работами (см. табл. 6).

ОБСУЖДЕНИЕ

В этом исследовании изучалось количество отозванных статей в год в базе данных WoS за 45-летний интервал, в частности, чтобы выяснить, существует ли взаимосвязь между квартилем журнала и количеством отозванных статей в журналах, перечисленных в WoS, а также существует ли какая-либо связь между отозванными статьями и статьями, которые они публикуют, и статьями, в которых они цитируются.

Согласно нашим выводам, за рассматриваемый период увеличилось количество отозванных статей. В настоящее время наблюдается значительное снижение после заметного пика в 2015г. Похоже,

что обнаружение отозванных статей в последнее время стало более эффективным благодаря развитию систем распознавания мошенничества и более тщательному рассмотрению в процессе рецензирования, а также, в частности, давлению со стороны коллег. В результате недействительная (лже)наука выявляется быстрее, чем когда-либо прежде [38]. Растущее число отозванных статей должно вызывать озабоченность из-за негативных последствий, которые оказывает такой отзыв; иногда могут быть опасные последствия из-за применения отозванных статей, в которых в реальный мир были анонсированы результаты, особенно в дисциплинах, связанных со здоровьем*. Этот аспект становится очень важным, особенно когда опровержение является неправомерным поведением или мошенничеством. К сожалению, но есть факт, что в последние годы огромное количество статей было отозвано из-за неправомерных действий в области исследований [39, 10, 40, 41].

Результаты настоящего исследования показывают, что наибольшее количество отозванных научных статей приходится на журналы с высоким импакт-фактором (Q1). Наши выводы согласуются с выводами Стинга [42], основанными на данных, извлеченных из базы данных PubMed. Аналогичным образом, Фанг и др. [39] также указали на значительную корреляцию между импакт-фактором журналов и мошенничеством или честными ошибками. Авторы [43] также показали, что журналы с высокими импакт-факторами обычно имеют больше отозванных статей, чем журналы с низкими импакт-факторами в области инженерии. Более того, мы наблюдали значительную корреляцию между квар-

* Случай потенциального действия гидроксихлорина в качестве лекарственного средства против Covid-19 известен во всем мире на момент написания этой статьи.

тиями журналов и количеством отозванных статей. В ходе расследования политики отзыва журналов с высоким импакт-фактором, индексируемых в базе данных JCR 1999, автор работы [44] сообщил, что 3,27% журналов (т.е. 4 из 122) придерживались политики отзыва. Таким образом, редакторам журналов было бы лучше добавить политику отзыва из публикации на веб-сайте журнала, чтобы проинформировать своих авторов. Более того, некоторые примеры отзыва могут быть полезны авторам, поскольку эти примеры могут четко показать, как их статьи могут быть отозваны.

Анализ второй гипотезы предполагает, что отозванные статьи входят в число цитирований значительного числа поданных статей. Это привело бы к вероятности того, что научная основа опубликованных статей была бы бессознательно и неосознанно выведена из ложных утверждений отозванных статей. Идентифицировать этот тип влияния сложнее, чем распознавать публикации и научные статьи, которые были отозваны [2]. Это особенно важно в таких чувствительных областях, как медицина, ветеринария и растениеводство, изучающих вопросы, связанные с жизнью. Хотя допустимо, чтобы исследователи использовали и ссылались на отозванные статьи как на неподходящие примеры и опасные методы распространения знаний, также рекомендуется, чтобы они не включали эти статьи в раздел библиографии своих статей. Вместо этого предлагается, чтобы ссылки на эти статьи содержались только в сносках [45, 46].

Анализ третьей гипотезы предполагает, что в отозванных статьях цитировались поданные статьи чаще, чем отозванные статьи. Казалось бы, отозванные статьи не должны были приводить к выпуску последующих отозванных статей. Особенности технической ответственности в этом выводе заключается в том, следует ли опускать ссылки в отозванных статьях, в которых цитировались поданные статьи, при измерении количества ссылок на поданные статьи или нет. С нашей точки зрения, базы данных цитирования, такие как WoS и Scopus, должны решить эту проблему. Действительно, когда одна статья отозвана, несправедливо рассчитывать количество цитируемых в ней ссылок как правильную цитируемость в рассчитанном количестве цитирований предыдущих статей, потому что это отозванная статья. Такие ссылки не должны учитываться при расчете, например, индекса Хирша [33, 36] или при измерении веса соавторов [47, 48].

Таким образом, эта тема усиливает озабоченность в обсуждении цитирования статей, которые были признаны надежными источниками. В качестве стратегии анализа проблемы цитирования отозванных публикаций авторы [49] утверждают, что программное обеспечение для управления ссылками должно осуществлять проверку отзыва во время онлайн подачи научных статей. Это позво-

ляет обнаружить любые отозванные статьи, которые авторы, возможно, неосознанно цитировали в своих статьях. Это может быть реализовано с помощью современных методов [50, 51]. Мы предполагаем, что количество цитирований поданных статей в отозванных статьях также следует опустить. Более того, авторы [52] и [49] предполагают, что статус отозванной статьи должен быть показан вместе с названием работы или выделен красочной полосой рядом с названием. Следует отметить, что не все уведомления об отзыве всегда являются прозрачными и четкими. Это приводит к тому, что отозванные статьи постоянно получают цитаты из других статей, включая поданные статьи. Поэтому рекомендуется, чтобы Комитет по этике публикаций (COPE) предложил уникальные инструкции по написанию уведомления об отзыве. Ван дер Вет и Найвин [53] также предполагают, что авторам, чьи статьи были выявлены за включение сокращенных цитат, должно быть разрешено, даже после публикации их статей, изменять свои статьи, удаляя отозванные ресурсы.

С другой стороны, эта процедура может уменьшить негативное влияние роста науки. Учитывая, что в отозванных статьях в значительной степени цитируются неизданные научные статьи, предлагается определять влияние предыдущих статей на последующие статьи. Если этот эффект основан на методе исследования или других существенных аспектах исследования, особенно в области наук о здоровье и медицины, впоследствии опубликованные статьи должны быть осведомлены и как можно быстрее очищаться от побочных эффектов, вызванных отзывом статьи.

В конце концов, следует отметить, что все отозванные статьи не относятся к одному и тому же типу и возникают по нескольким причинам, рассмотренным выше [6,12]. Следовательно, не следует принимать одно и то же решение в отношении всех отозванных статей. Например, отозванная статья из-за манипуляций с цифрами все еще может считаться обоснованной, поскольку ее методология остается тщательной [49]. Придерживаясь этой точки зрения, владельцам баз данных цитирования рекомендуется разработать индекс отзыва, который вычисляет вес отзыва для отозванных статей на основе их типа отзыва. Такой индекс предполагает порог, выше которого статьи никоим образом не должны использоваться. Кроме того, учитывая, что многие авторы могут быть не знакомы со случаями отзыва, рекомендуется, чтобы журналы размещали примеры отозванных статей на своих веб-сайтах. Многие журналы в разделе, который у них есть в качестве памятки для авторов, отсылают авторов к руководствам по стилю цитирования, которые они используют, и полагают, что авторы тщательно следуют этим памяткам. Это полезно только для плагиата.

ЗАКЛЮЧЕНИЕ

Во многих поданных статьях цитировались отозванные статьи. Это не очень хорошие новости из-за возможной опасности, которую может вызвать отозванная статья. Таким образом, авторы поданных статей должны оценить свои статьи и сообщить редакторам, оказали ли они заметное влияние на их результаты. Журналы также могут выбрать более тщательное рецензирование статей, если у авторов были ранее отозванные статьи. Для этого журналы могут создавать список авторов отозванных статей, с которыми сравниваются авторы новых материалов перед процессом рецензирования. Это может помочь выявить мошенническую статью за более короткое время, поскольку, в соответствии с базой данных PubMed, довольно много отозванных статей, примерно 53% мошеннических статей, были написаны первым автором, который также написал другие отозванные статьи [42]. Также было бы интересно понаблюдать за страной [54] и/или институциональной принадлежностью авторов, отзывающих свои статьи [55].

По-прежнему необходимо провести глубокое исследование, чтобы понять, является ли причиной отзыва статьи использование ранее отозванных статей или допущение ошибок, таких как плагиат и манипулирование данными, независимо от типа неправомерного поведения [56], помимо ошибки издателя или автора(ов). Представляется крайне важным переоценить обоснованность утверждений, сделанных в поданных статьях, на которые ссылаются отозванные статьи, на случай, если их потребуется отозвать как можно скорее, чтобы они не повлияли на какую-либо будущую статью. В результате предлагается, чтобы базы данных обязывали журналы отслеживать статьи, которые используют отозванные статьи в качестве аргументов для своего отчета.

ЛИТЕРАТУРА

1. Shuai X., Rollins J., Moulinier I., Custis T., Edmunds M., Schilder F. A Multidimensional investigation of the effects of publication retraction on scholarly impact // *Journal of the Association for Information Science and Technology*. — 2017. — Vol. 68, No. 9. — P. 2225-2236. — <https://doi.org/10.1002/asi.23826>
2. Chen C., Hu Z., Milbank J., Schultz T. A visual analytic study of retracted articles in scientific literature // *Journal of the American Society for Information Science and Technology*. — 2013. — Vol. 64, No. 2. — P. 234-253. — <https://doi.org/10.1002/asi.22755>
3. Hertel C., Ausloos M., Ileanu B. V., Rotundo G., Andrei T. Quantitative and qualitative analysis of editor behavior through potentially coercive citations // *Publications*. — 2017. — Vol. 5, No. 2. — P. 15. — <https://doi.org/10.3390/publications5020015>
4. Sharma B. B., Singh V. Ethics in writing: Learning to stay away from plagiarism and scientific miscon-

duct // *Lung India: Official Organ of Indian Chest Society*. — 2011. — Vol. 28, No. 2. — P. 148-150. — <https://doi.org/10.4103/0970-2113.80337>

5. He T. Retraction of global scientific publications from 2001 to 2010 // *Scientometrics*. — 2013. — Vol. 96, No. 2. — P. 555-561. — <https://doi.org/10.1007/s11192-012-0906-3>

6. Budd J. M., Sievert M. E., Schultz T. R. Phenomena of retraction. Reasons for retraction and citations to the publications // *Jama-Journal of the American Medical Association*. — 1998. — Vol. 280, No. 3. — P. 296-297. — <https://doi.org/10.1001/jama.280.3.296>

7. Foo J. Y. A. A retrospective analysis of the trend of retracted publications in the field of Biomedical and Life sciences // *Science and Engineering Ethics*. — 2011. — 17, No. 3. — P. 459-468. — <https://doi.org/10.1007/s11948-010-9212-8>

8. Steen R. G. Retractions in the scientific literature: Do authors deliberately commit research fraud? // *Journal of Medical Ethics*. — 2011. — Vol. 37, No. 2. — P. 113-117. — <https://doi.org/10.1136/jme.2010.038125>

9. Grieneisen M. L., Zhang M. A comprehensive survey of retracted articles from the scholarly literature // *PLoS ONE*. — 2012. — Vol. 7, No. 10, e44118. — <https://doi.org/10.1371/journal.pone.0044118>

10. Marcus A., Oransky I. What studies of retractions tell us? // *Journal of Microbiology and Biology Education*. — 2014. — Vol. 15, No. 2. — <https://doi.org/10.1128/jmbe.v15i2.855>

11. Moylan E. C., Kowalczyk M. K. Why articles are retracted: A retrospective cross-sectional study of retraction notices at BioMed Central // *BMJ open*. — 2016. — Vol. 6, No. 11, e012047. — <http://dx.doi.org/10.1136/bmjopen-2016-012047>

12. Dal-Ré R., Ayuso C. Reasons for and time to retraction of genetics articles published between 1970 and 2018. // *Journal of Medical Genetics*. — 2019. — Vol. 56. — P. 734-740. — <http://dx.doi.org/10.1136/jmedgenet-2019-106137corr1>

13. Tchao R. A need to archive correct biomedical scientific data and to prevent continued citation of retracted scientific publications // *International Journal of Humanities and Arts Computing*. — 2014. — Vol. 8. — P. 29-37. — <https://doi.org/10.3366/ijhac.2014.0097>

14. Brainard J. Rethinking retractions // *Science*. — 2018. — Vol. 362, No. 6413. — P. 390-393. — <https://doi.org/10.1126/science.362.6413.390>

15. McCook A. One publisher, more than 7000 retractions // *Science*. — 2018. — Vol. 362, No. 6413. — P. 393. — <https://doi.org/10.1126/science.362.6413.393>

16. Al-Ghareeb A., Hillel S., McKenna L., Cleary M., Visentin D., Jones M., Bressington D., Gray R. Retraction of publications in nursing and midwifery research: A systematic review // *International Journal of Nursing Studies*. — 2018. — Vol. 81. — P. 8-13. — <https://doi.org/10.1016/j.ijnurstu.2018.01.013>

17. Fanelli D. How many scientists fabricate and falsify research? A systematic review and meta-analysis of

- survey data // PLoS ONE. — 2009. — Vol. 4, No. 5, e5738. — <https://doi.org/10.1371/journal.pone.0005738>
18. Steen R. G., Casadevall A., Fang F. C. Why has the number of scientific retractions increased? // PLoS ONE. — 2013. — Vol. 8, No. 7, e68397. — <https://doi.org/10.1371/journal.pone.0068397>
19. Fang F. C., Casadevall A. Retracted science and the retraction index // Infection and Immunity. — 2011. — Vol. 79, No. 10. — P. 3855-3859. — <https://doi.org/10.1128/iai.05661-11>
20. Stern A. M., Casadevall A., Steen R. G., Fang F. C. Financial costs and personal consequences of research misconduct resulting in retracted publications // Elife. — 2014. — Vol. 3, e02956. — <https://doi.org/10.7554/eLife.02956>
21. Wager E., Williams P. Why and how do journals retract articles? An analysis of Medline retractions 1988–2008. // Journal of Medical Ethics. — 2011. — Vol. 37, No. 9. — P. 567-570. — <http://dx.doi.org/10.1136/jme.2010.040964>
22. Decullier E., Huot L., Samson G., Maisonneuve H. Visibility of retractions: A cross-sectional one-year study // BMC Research Notes. — 2013. — Vol. 6, 238. — <https://doi.org/10.1186/1756-0500-6-238>
23. Furman J. L., Jensen K., Murray F. Governing knowledge in the scientific community: Exploring the role of retractions in biomedicine // Research Policy. — 2012. — Vol. 41, No. 2. — P. 276-290. — <https://doi.org/10.1016/j.respol.2011.11.001>
24. Nath S. B., Marcus S. C., Druss B. G. Retractions in the research literature: Misconduct or mistakes? // Medical Journal of Australia. — 2006. — Vol. 185, No. 3. — P. 152-154. — <https://doi.org/10.5694/j.1326-5377.2006.tb00504.x>
25. Trikalinos N. A., Evangelou E., Ioannidis J. P. A. Falsified papers in high-impact journals were slow to retract and indistinguishable from nonfraudulent papers // Journal of Clinical Epidemiology. — 2008. 61, No. 5. — P. 464-470. — <https://doi.org/10.1016/j.jclinepi.2007.11.019>
26. Bar-Ilan J., Halevi G. Post retraction citations in context: A case study // Scientometrics. — 2017. — Vol. 113, No. 1. — P. 547-565. — <https://doi.org/10.1007/s11192-017-2242-0>
27. Budd J. M., Sievert M., Schultz T. R., Scoville C. Effects of article retraction on citation and practice in medicine // Bulletin of the Medical Library Association. — 1999. — Vol. 87, No. 4. — P. 437-443. — <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC226618/>
28. Korpela K. M. How long does it take for the scientific literature to purge itself of fraudulent material? The Breuning case revisited // Current Medical Research and Opinion. — 2010. — Vol. 26, No. 4. — P. 843-847. — <https://doi.org/10.1185/03007991003603804>
29. Steen R. G. Retractions in the scientific literature: Is the incidence of research fraud increasing? // Journal of Medical Ethics. — 2011. — Vol. 37, No. 4. — P. 249-253. — <https://doi.org/10.1136/jme.2010.040923>
30. Pfeifer M. P., Snodgrass G. L. The continued use of retracted, invalid scientific literature // Jama-Journal of the American Medical Association. — 1990. — Vol. 263, No. 10. — P. 1420-1423.
31. Campanario J. M. Fraud: Retracted articles are still being cited // Nature. — 2000. — 408, No. 6810. — P. 288-288. — <https://doi.org/10.1038/35042753>
32. Couzin J., Unger K. Cleaning up the paper trail // Science. — 2006. — Vol. 312, No. 5770. — P. 38-43. — <https://doi.org/10.1126/science.312.5770.38>
33. Teixeira da Silva J. A., Dobránszki J. Notices and policies for retractions, expressions of concern, errata and corrigenda: Their importance, content, and context // Science and Engineering Ethics. — 2017. — Vol. 23, No. 2. — P. 521-554. — <https://doi.org/10.1007/s11948-016-9769-y>
34. Hamilton D. G. Continued citation of retracted radiation oncology literature – Do we have a problem? // International Journal of Radiation Oncology, Biology, Physics. — 2019. — Vol. 103, No. 5. — P. 1036-1042. — <https://doi.org/10.1016/j.ijrobp.2018.11.014>
35. Rubbo P., Pilatti L. A., Picinin C. T. Citation of retracted articles in engineering: A study of the web of science database // Ethics & Behavior. — 2019. — Vol. 29, No. 8. — P. 661-679. — <https://doi.org/10.1080/10508422.2018.1559064>
36. Teixeira da Silva J. A., Dobránszki J. Highly cited retracted papers // Scientometrics. — 2017. — Vol. 110, No. 3. — P. 1653-1661. — <https://doi.org/10.1007/s11192-016-2227-4>
37. Graczyk M. B., Duarte Queirós S. M. Intraday seasonalities and nonstationarity of trading volume in financial markets: Collective features // PLoS ONE. — 2017. — Vol. 12, No. 7, e0179198. — <https://doi.org/10.1371/journal.pone.0179198>
38. Bornemann-Cimenti H., Szilagyi I. S., Sandner-Kiesling A. Perpetuation of retracted publications using the example of the Scott s. Reuben Case: Incidences, reasons and possible improvements // Science and Engineering Ethics. — 2016. — Vol. 22, No. 4. — P. 1063-1072. — <https://doi.org/10.1007/s11948-015-9680-y>
39. Fang F. C., Steen R. G., Casadevall A. Misconduct accounts for the majority of retracted scientific publications // Proceedings of the National Academy of Sciences of the United States of America. — 2012. — Vol. 109, No. 42. — P. 17028-17033. — <https://doi.org/10.1073/pnas.1212247109>
40. Samp J. C., Schumock G. T., Pickard A. S. Retracted publications in the drug literature // Pharmacotherapy. — 2012. — Vol. 32, No. 7. — P. 586-595. — <https://doi.org/10.1002/j.1875-9114.2012.01100.x>
41. van Noorden R. Science publishing: The trouble with retractions // Nature. — 2011. — Vol. 478, No. 7367. — P. 26-28. — <https://doi.org/10.1038/478026a>
42. Steen R. G. Retractions in the medical literature: How many patients are put at risk by flawed research? // Journal of Medical Ethics. — 2011. —

Vol. 37, No. 11. — P. 688-692. — <https://doi.org/10.1136/jme.2011.043133>

43. *Rubbo P., Helmann C. L., Bilynskiyycz dos Santos C., Pilatti L. A.* Retractions in the Engineering Field: A study on the Web of Science database // *Ethics & Behaviour*. — 2019. — Vol. 29, No. 2. — P. 141-155. — <https://doi.org/10.1080/10508422.2017.1390667>

44. *Atlas M. C.* Retraction policies of high-impact biomedical journals. // *Journal of the Medical Library Association*. — 2004. — Vol. 92, No. 2. — P. 242-250. — <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC385306/>

45. *Eisenach J. C.* Data fabrication and article retraction how not to get lost in the woods // *Anesthesiology: The Journal of the American Society of Anesthesiologists*. — 2009. — Vol. 110. — P. 955-956. — <https://doi.org/10.1097/ALN.0b013e3181a06bf9>

46. *Cosentino A. M., Veríssimo D.* Ending the citation of retracted papers // *Conservation Biology*. — 2016. — Vol. 30, No. 3. — P. 676-678. — <https://doi.org/10.1111/cobi.12676>

47. *Ausloos M.* Assessing the true role of coauthors in the h-index measure of an author scientific impact // *Physica A: Statistical Mechanics and its Applications*. — 2015. — Vol. 422, 136-142. — <https://doi.org/10.1016/j.physa.2014.12.004>

48. *Ausloos M.* Coherent measures of the impact of co-authors in peer review journals and in proceedings publications // *Physica A: Statistical Mechanics and its Applications*. — 2015. — Vol. 438, 568-578. — <https://doi.org/10.1016/j.physa.2015.06.010>

49. *Teixeira da Silva J. A., Bornemann-Cimenti H.* Why do some retracted papers continue to be cited? // *Scientometrics*. — 2017. — Vol. 110, No. 1. — P. 365-370. — <https://doi.org/10.1007/s11192-016-2178-9>

50. *Mrowinski M. J., Fronczak A., Fronczak P., Nedic O., Ausloos M.* Review time in peer review: Quantitative

analysis and modeling of editorial workflows // *Scientometrics*. — 2016. — Vol. 107, No. 1. — P. 271-286. — <https://doi.org/10.1007/s11192-016-1871-z>

51. *Mrowinski M. J., Fronczak P., Fronczak A., Ausloos M., Nedic O.* Artificial intelligence in peer review: How can evolutionary computation support journal editors? // *PloS one*. — 2017. — Vol. 12, No. 9, e0184711. — <https://doi.org/10.1371/journal.pone.0184711>

52. *Oransky I., Marcus A.* Why write a blog about retractions? — [Retraction Watch Blog, 2010, 2 August]. — <http://retractionwatch.com/2010/08/03/why-write-a-blog-about-retractions/>

53. *van der Vet P. E., Nijveen H.* Propagation of errors in citation networks: A study involving the entire citation network of a widely cited paper published in, and later retracted from, the journal *Nature* // *Research Integrity and Peer Review*. — 2016. — Vol. 1, 3. — <https://doi.org/10.1186/s41073-016-0008-5>

54. *van Leeuwen T. N., Lamel M.* An in-depth analysis of papers retracted in the Web of Science/ Context Counts: Pathways to Master Big and Little Data // *Proceedings of the science and technology indicators conference* (pp. 337-344). — 2014.

55. *Ahmed K.* Academic integrity: Challenges and strategies for Asia and the Middle East // *Accountability in Research*. — 2020. — Vol. 27, No. 5. — P. 256-270. — <https://doi.org/10.1080/08989621.2019.1646646>

56. *DuBois J. M., Anderson E. E., Chibnall J., Carroll K., Rubbelke T.* Understanding research misconduct: A comparative analysis of 120 cases of professional wrongdoing // *Accountability in Research*. — 2013. — Vol. 20. — P. 320-338. — <https://doi.org/10.1080/08989621.2013.822248>

Приглашаем российских и зарубежных авторов к сотрудничеству
в журнале «Международный форум по информации».
Оригинальные статьи и другие материалы (рецензии, письма)
можно присылать на русском или английском языке
по почтовому адресу, указанному в «Памятке для авторов»
или по электронной почте: mfi@viniti.ru.

Ответственный за выпуск *Е.А. Овченкова*

Компьютерная верстка *М. А. Филимонова*

ИД № 04689 от 28.04.2001 г.

Подписано в печать 07.06.2022 г.

Бумага офсетная. Формат 60х84 1/8. Гарн. литер. Печать цифровая

Усл. печ. л. 4,00 Уч.-изд. л. 4,30 Тираж 33 экз.

Адрес редакции: 125190, Россия, г. Москва, ул. Усиевича, д. 20.

Тел. (499) 155-44-95

Отпечатано в ООО «Типография Форпринт».

г. Москва, ул. Маршала Бирюзова, д. 1, корп. 11. Тел. +7 (495) 211-38-33

ВИНИТИ РАН

Центр научно-информационного обслуживания

Информационные услуги, предоставляемые ЦНИО ВИНТИ РАН:

- проведение тематического поиска и консультации поисковых экспертов;
- подготовка списков научной литературы;
- подбор, копирование полнотекстовых материалов из первоисточников на бумажном носителе и в электронном виде;
- библиометрическая оценка публикационной активности исследователей и научных организаций с использованием российских и зарубежных баз данных;
- информационное обеспечение информационно-аналитической деятельности по подготовке и предоставлению аналитических обзоров и других научных материалов.

ВИНИТИ РАН располагает следующими информационными ресурсами:

- фондом НТЛ, включающим более 2,5 млн. отечественных и иностранных журналов, книг, депонированных рукописей, авторефератов диссертаций и другой научной литературы, ретроспектива – с 1991 года;
- базами данных и Интернет-ресурсами: БД ВИНТИ (разработка ВИНТИ), БД SCOPUS, БД Questel (патенты) и другими реферативными ресурсами;
- полнотекстовыми электронными ресурсами (статьи, патенты, материалы конференций).

Ознакомиться с информацией о доступных полнотекстовых и реферативных ресурсах можно на сайте ВИНТИ РАН www.viniti.ru

К услугам пользователей – **Электронный Каталог ВИНТИ** <http://catalog.viniti.ru>
и служба электронной доставки документов.

Осуществляется платное информационное обслуживание по разовым заказам и на договорной основе с предоставлением всех необходимых финансовых документов.

Проводится индивидуальное обслуживание пользователей в читальном зале ЦНИО ВИНТИ РАН.

Подробную информацию Вы можете получить:

Адрес: 125190, Россия, г. Москва, ул. Усиевича, 20, ВИНТИ РАН;

Телефоны: 499-155-42-17, 499-155-42-43;

E-mail: cnio@viniti.ru

УВАЖАЕМЫЕ КОЛЛЕГИ!

ВИНИТИ РАН предлагает Вашему вниманию Реферативный Журнал в электронной форме

РЖ в электронной форме (ЭлРЖ) выпускается по всем разделам естественных, технических и точных наук.

Каждый номер ЭлРЖ является полным аналогом печатного номера РЖ по составу описаний документов, их оформлению и расположению. Он сопровождается оглавлением, указателями.

ЭлРЖ представляет собой информационную систему, снабженную поисковым аппаратом и позволяющую пользователю на персональном компьютере:

- читать номер РЖ, последовательно листая рефераты;
- просматривать рефераты отдельных разделов по оглавлению;
- обращаться к рефератам по указателям авторов, источников, ключевых слов;
- проводить поиск документов по словам и словосочетаниям;
- выводить текст описаний документов во внешний файл.

ЭлРЖ могут быть:

- записаны на DVD-ROM;
- передаваться через FTP-сервер (клиенту предоставляется логин и пароль с доступом к FTP-серверу ВИНТИ, с которого он скачивает заказанные журналы).

Электронные реферативные журналы можно заказать за текущий год с любого номера, а также за предыдущие годы.

Подробную информацию Вы можете получить:

Адрес: 125190, Россия, Москва, ул. Усиевича, 20, ВИНТИ РАН

Телефон: 8 499-152-62-11

E-mail: feo@viniti.ru