МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ

ВСЕРОССИЙСКИЙ ИНСТИТУТ НАУЧНОЙ И ТЕХНИЧЕСКОЙ ИНФОРМАЦИИ
РОССИЙСКОЙ АКАДЕМИИ НАУК
(ВИНИТИ РАН)

НАУЧНО • ТЕХНИЧЕСКАЯ ИНФОРМАЦИЯ

Серия 2. ИНФОРМАЦИОННЫЕ ПРОЦЕССЫ И СИСТЕМЫ

ЕЖЕМЕСЯЧНЫЙ НАУЧНО-ТЕХНИЧЕСКИЙ СБОРНИК

Издается с 1961 г. **№ 11** **Москва 2020**

ОБЩИЙ РАЗДЕЛ

УДК 004.6:004.774

Е.Ю. Дмитриева, О.В. Сюнтюренко

Актуальные задачи диверсификации технологий, информационных продуктов и услуг

Рассматриваются объективные факторы-предпосылки корректировки концептуальных положений и задач информационного обеспечения современной научно-промышленной сферы. Показан спектр актуальных и перспективных направлений создания и внедрения новых технологий, информационных продуктов и услуг в структуре Государственной системы НТИ. Представлен вариант методологического подхода к управлению информационным обеспечением на основе оценки уровня информационной поддержки целевых комплексных программ и проектов. Структурирован потенциальный экономический макроэффект, определяемый результативностью решения комплекса задач по повышению уровня информационной поддержки исследований, разработок, трансфера технологий.

Ключевые слова: *информационные продукты, конвергенция технологий, интернет-ресурсы, информационная инфраструктура, аналитическая постобработка, навигация и поиск, реферативная информация, базы данных, трансфер технологий*

DOI: 10.36535/0548-0027-2020-11-1

ВВЕДЕНИЕ

Задачи повышения эффективности информационной поддержки исследований, разработок, трансфера технологий чрезвычайно актуализируются в современных сложных условиях российской экономики и стимулируют поиск новых подходов и инновационных решений. Масштаб и размах происходящих изменений обусловливают рост дизруптивных инноваций (инноваций, которые изменяют соотношение ценностей на рынке) практически во всех сегментах научно-промышленной сферы. Темпы развития и распространения инноваций оказываются беспрецедентно быстрыми. Таким образом, динамичность информационной среды, развитие и внедрение новейших технологий остро ставит вопрос о формировании новых информационных продуктов, услуг и использовании новых технологий информационного обеспечения. Особую значимость и актуальность приобретает производство информационно-аналитических продуктов и услуг для поддержки развития высокотехнологичных секторов экономики, технико-экономического анализа объектов и процессов в различных разрезах, многоаспектного моделирования и прогнозирования. Формирование цифровых информационных ресурсов и становление информационных технологий (как инструментов моделирования и продуцирования новых технологий и услуг) открывают новые приложения и возможности эффективного информационного обеспечения, моделирования, прогнозирования и планирования.

Существенную корректировку концептуальных положений и задач информационного обеспечения научно-промышленной сферы определяют следующие объективные факторы:

- доминирующий тренд новой информационной среды, что проявляется в быстром росте объема цифровых данных, интернет-ресурсов и перманентном расширении глобальной сети телекоммуникаций. Развитие глобальной сети Интернет влечет смену парадигмы функционирования системы информационного обеспечения исследований и разработок – от иерархической к сетевой. Создание распределенных сетевых информационных ресурсов (ИР) является наиболее бурно развивающимся направлением информатизации научно-промышленной сферы;
- конвергенция информационных, традиционных библиотечных, компьютерных и телекоммуникационных технологий. Самоорганизация (в смысле адаптивности структуры и функциональных ролей участников) глобальной сетевой институциональной среды;
- виртуальная и реальная интеграция разнородных информационных ресурсов в гетерогенной цифровой среде. Преимущественное использование интероперабельных программных средств, унифицированных и отчуждаемых программных и технологических решений, современных сервисно-ориентированных архитектур, специализированных систем навигации к проблемно-ориентированным информационным ресурсам;
- растущая востребованность развития современных методов и средств извлечения и объединения знаний из различных типов информационных источников с последующим наложением на них различных связей между исследуемыми объектами. Здесь значительные перспективы имеет мультипликативная аналитическая постобработка научно-технической и технико-экономической информации с использованием методов наукометрии, эконометрии и многомерного анализа данных;
- актуализация задач управления знаниями и информационной поддержки принятия решений. В прагматическом аспекте – это пакет технологий, включающий в себя комплекс формализованных методов, охватывающих: поиск и извлечение знаний; структурирование и систематизацию знаний; анализ знаний (выявление зависимостей и аналогий); обновление (актуализацию) знаний; распространение знаний; генерацию новых знаний.

Эти факторы и актуальные задачи развития российской экономики требуют создания качественно новых технологий информационной поддержки научных исследований и наукоемкого производства как ключевого фактора ускоренного технологического развития. Далее мы рассмотрим спектр актуальных и перспективных направлений создания и внедрения технологий, информационных продуктов и услуг в структуре Государственной системы НТИ и, прежде всего, в ВИНИТИ РАН как ее ведущей организации.

НОВЫЕ ПЕРСПЕКТИВНЫЕ НАПРАВЛЕНИЯ СОЗДАНИЯ ИНФОРМАЦИОННЫХ ПРОДУКТОВ И УСЛУГ

В прагматическом аспекте в современной научно-промышленной сфере необходимость создания новых методов обработки, а также продуктов и услуг предопределяет устойчивые тенденции временного сжатия цикла «исследование – разработка – производство» и роста удельного веса исследований и разработок междисциплинарного характера. Интеграция информационных технологий с индустриальными коренным образом перестраивает процессы производства и актуализирует создание новых инструментов и методов информационной поддержки.

С системных позиций на содержательном уровне кратко рассмотрим континиум {A,B,C,D,E,F,G,H,I,K,L} перспективных методов, продуктов и услуг, реализация которых позволит на качественно более высоком уровне решать задачи информационного обеспечения исследований и разработок.

A. ***Разработка системы навигации и поиска знаний в гетерогенной сетевой среде на основе универсального интеллектуального конвертора метаданных.*** В современных условиях, характеризуемых лавинообразным ростом объемов научной информации, разнообразием ее видов и форм представления, задача поиска информации критически усложняется. Сегодня теория научно-технической информации не располагает методами индустриальной интеграции знаний, представленных в разнородных источниках. В мировом информационном пространстве основным видом поиска научной информации является поиск по свободной лексике (лексический поиск), на котором основаны распространенные поисковые машины (Яндекс, Google). Однако такой поиск дает низкие

характеристики полноты и точности, в частности, потому, что при нем не учитываются семантические связи понятий. В ВИНИТИ РАН создана методология и ведется разработка системы, обеспечивающей эффективный поиск информации в пространстве разнородных ресурсов, содержащих данные, проиндексированные по различным системам классификации, ключевым словам, средствам полнотекстового поиска. Этот проект, поддержанный грантами РФФИ (№ 17-07-00153 и № 20-07-00103), включает разработку и реализацию алгоритмов автоматического конвертирования поисковых запросов, поступающих на естественном языке, в форму, обеспечивающую поиск информации с использованием различных классификационных языков. Онтология пространства научных знаний может быть представлена как сеть семантических связей понятий, отображаемых ключевыми словами и классификационными рубриками. Специализированная база данных, поддерживающая разработанную онтологию, будет служить основой для смысловой навигации по источникам, структурированным различными системами индексирования. Это позволит обеспечить эффективный поиск научной информации в сетевых условиях разнородности информационных ресурсов, что является исключительно важной задачей в современных условиях [1-3].

B. ***Реализация тематико- и/или проблемно-ориентированного избирательного распространения информации (ИРИ) с использованием интернет-СМИ и научных журналов открытого доступа.*** Потенциал и актуальность этого направления определяются двумя тренд-факторами.

Первое. Рост интернет-ресурсов в последние десять лет приобретает лавинообразный характер. По данным International Data Corporation (IDC), мировой объем информации удваивается каждые два года. По некоторым оценкам суммарный объем всех научных журналов в мире за один год составляет более 1 Тб информации, которая из-за огромных объемов практически необозрима в открытом доступе. Поиск нужной информации в Интернете становится все более важной и сложной задачей. Можно с уверенностью утверждать, что потенциально любой специалист мог бы найти в СМИ достаточно много интересной, новой и актуальной информации научно-технического, правового, финансового и экономического характера.

Второе. В последнее десятилетие развитые страны обращают все большее внимание на проблему открытости и доступности результатов научных исследований. В перспективе это позволит повысить прозрачность науки, сократить нерациональные затраты, существенно снизить издержки финансирования дублирующих исследований. По плану Европейской комиссии в рамках 7-летней программы Horizon 2020 (бюджет $80 млрд), более 80% всех публикаций европейских учёных, проводящих свои исследования за государственный счёт, будут размещаться в журналах открытого доступа. Некоторые страны (Австралия, Великобритания, США, и др.) уже сейчас на самом высоком уровне занимаются решением этой проблемы [4]. В России реализуется пилотный проект КиберЛенинка, где по лицензионному договору размещаются научные журналы открытого доступа. В списке имеющихся в eLibrary журналов открытого доступа свыше 3-х тыс. наименований.

Развитие интернет-ИРИ как новой системы информационного обслуживания должно базироваться на использовании механизма кластеризации потоковой информации из открытых источников с использованием методов построения адаптивных гипермедиа на основе технологии кластеризации неструктурированных данных и обеспечения способа донесения актуальной, лингвистически обработанной информации до различных целевых групп ее потребления (и отдельных пользователей) в соответствии с их персональными потребностями и ожиданиями. С некоторой долей условности можно говорить о создании ИРИ (избирательного распространения информации) нового поколения на основе конвергенции телекоммуникационных, компьютерных и информационных технологий. Качественно новый уровень конвергированного ИРИ характеризуется: практически неограниченным кругом источников (и пользователей), предельной минимизацией временного лага, высокой целевой избирательностью. При реализации информационного комплекса должны быть использованы методы вычислительной математики и компьютерной лингвистики, предназначенные для обработки текста на естественном языке, такие как вероятностный морфологический анализ, синтаксический анализ и ранжирование, синтаксический анализ и эксплицирование отношений, установление референтных связей и др. В полном объеме реализацию ИРИ нового поколения ВИНИТИ РАН мог бы осуществить в достаточно сжатые сроки [4].

C. ***Реализация информационного обслуживания на основе политематического полнотекстового банка данных и федерального индекса научного цитирования.*** Политематический полнотекстовый банк данных ВИНИТИ РАН помимо зарубежных изданий должен содержать все отечественные научно-технические журналы. Для реализации этого направления следует выполнить значительный объем подготовительных работ (создание аппаратной платформы, выбор и инсталляция программного обеспечения, разработка договорной базы и правовых вопросов, проведение стоимостной оптимизации и др.). Необходимое условие для этого – первоочередное решение задач комплектования входного потока научной литературы.

В ВИНИТИ РАН функционирует оригинальная система автоматизированной обработки входного потока литературы, обеспечивающая взаимодействие десятков операторов на всех стадиях технологического процесса. Система охватывает все виды изданий: журналы, книги, депонированные рукописи, описания изобретений (патентную литературу), стандарты, материалы конференций и пр. Сформированная производственная база позволяет сканировать первоисточники, поступающие как по подписке, так и из других источников (библиотеки, личные экземпляры и др.). В настоящее время эта система в основном используется для обеспечения научных отделов Института копиями статей, необходимых для создания Реферативного журнала (РЖ) и полнотекстовой Техно-

логической БД первоисточников, поступающих в ВИНИТИ. На имеющемся оборудовании обрабатывается свыше 0,5 млн статей в год (свыше 2 млн страниц).

Производственный процесс включает этапы постатейной регистрации первоисточников, сканирования и распечатки результатов сканирования, а также полного системного программного сопровождения этих этапов. Система решает задачи учета поступающих экземпляров, регистрации выпусков изданий, аналитической обработки выпусков (постатейно) и тематической разметки выделенных документов. Критической позицией является разработка и/или адаптация надежно функционирующей биллинговой[1] системы и организация системы взаимных расчетов с издательствами и другими поставщиками данных.

Один из важных ресурсов информатизации науки индекс научного цитирования – это принятая в научном мире мера "значимости" научной работы. Единственным источником сведений в индексах цитирования является только сама публикация, а политика взаимодействия индексов с государством и пользователями должна быть гибкой. Российский индекс научного цитирования (РИНЦ) не является государственным в полном смысле, а принципы его формирования, не предусматривающие критериев отбора изданий и предоставляющие организациям и авторам самим корректировать свои данные, создают сложности при его использовании для решения государственных задач учета и оценки результатов научной деятельности. Необходимо разработать индекс научного цитирования, отвечающий требованиям отбора и представления необходимой наукометрической информации на более высоком качественном уровне и, главное, являющийся государственным ресурсом. Наличие в ВИНИТИ РАН реферативного информационного ресурса большого объема в электронном виде и фонда первоисточников – русскоязычных журналов и книг, позволяет достаточно оперативно и на хорошем уровне начать создание информационного ресурса нового типа – федерального (государственного) индекса научного цитирования [5]. Банк данных ВИНИТИ (более 36 млн док.) значительно сокращает объем работ по описанию включаемых в индекс публикаций, однако не заменяет библиографическую обработку сносок и пристатейных списков литературы – самой затратной части создания ресурса. Кроме того, банк данных включает далеко не все необходимые публикации из выбранных изданий, поэтому дополнение его ретроспективной информацией также придется предусмотреть, как и решение многих других вопросов. Было бы целесообразно рассмотреть и возможность использования разработок *Thomson-Reuter* и *Elsevier* по созданию индексов цитирования *Web of Science* и *Scopus*. Это в дальнейшем позволит интегрировать и оперативно дополнить эти мировые базы данных сведениями о российских изданий.

[1] Автоматизированная система расчетов, ответственных за сбор информации об использовании телекоммуникационных услуг, их тарификацию, выставление счетов абонентам, обработку платежей.

D. ***Реализация практического применения новых методов и систем автоматического перевода научно-технических текстов.*** Задача минимизации негативных факторов, обусловленных языковым «барьером» и детерминирующих использование электронных информационных ресурсов, довольно успешно решается путем более широкого внедрения в практику поиска и обработки информации систем автоматического и автоматизированного перевода. Действующие системы машинного перевода ориентированы на конкретные пары языков (например, английский и русский). Качество машинного перевода зависит от объема словаря, объема информации, приписываемой лексическим единицам, от тщательности составления и проверки алгоритмов анализа и синтеза, от эффективности программного обеспечения.

В последнее десятилетие стал доминировать статистический подход к машинному переводу. Параметры статистических моделей, на основе которых генерируется перевод, являются производными от анализа двуязычных корпусов текста (*text corpora*). Компьютеры оценивают статистические закономерности в больших массивах ранее накопленного цифрового контента. Самообучение компьютера осуществляется посредством анализа достаточно большого (сотни тысяч) количества параллельных текстов – содержащих одинаковую информацию на разных языках. Например, Евросоюз и ООН выпускают множество текстов документов на всех основных языках стран-участниц [6]. Основным преимуществом статистических систем является их свойство не отставать от развития и подвижности языка: если в языке происходят какие-либо изменения, то система сразу это распознает и самостоятельно обучается, при этом качественно перевод отличается гладкостью [7, 8]. Периодические издания ВИНИТИ на русском и английском языках, огромный объем реферативного БнД (более 36 млн документов, с глубиной ретроспективы по некоторым предметным областям до 15 лет), значительный входной поток информации – всё это создает реальные предпосылки для применения новых статистических методов машинного перевода с целью повышения эффективности информационного обслуживания российских и зарубежных ученых и специалистов.

E. ***Производство информационно-аналитических продуктов и услуг с использованием методов наукометрии и многомерного анализа данных.*** Мультипликативная аналитическая постобработка научно-технической и технико-экономической информации с использованием методов наукометрии и многомерного анализа данных позволяет выявлять статистические закономерности, выражающие зависимости между распределениями различных параметров исследуемых систем и процессов, и характер изменения распределений во времени [9-12]. Использование методов аналитической постобработки реферативной и библиографической информации представляется весьма перспективным для решения ряда задач, в числе которых:

- анализ структуры отечественной и мировой науки;
- определение тенденций и процессов, происходящих в мировой и региональной науке;

o выявление наиболее актуальных или, напротив, теряющих свою актуальность научных направлений;

o отслеживание генезиса конкретных научных идей и истории их развития;

o определение продуктивности работы исследователей в конкретной научной области и эффективности материальных затрат в этой области;

o анализ структуры научного сообщества и изучение науки как социального организма.

Исходной ресурсной базой, помимо реферативного БнД ВИНИТИ, могут быть и ресурсы БнД Российского фонда фундаментальных исследований (*www.rfbr.ru*), Федеральных научно-технических программ (*www.fcntp.ru*), Росстата (*www.gks.ru*) и др. Совместная постобработка информации БнД ВИНИТИ и данных Росстата (ВВП, произведенной энергии, среднего годового дохода на душу населения, произведенного продукта с использованием высоких технологий и ряда других) – это перспективное множество представляющих практический интерес статистических показателей и распределений, позволяющих анализировать:

- сравнительный рост ВВП, расходов на образование, исследования и разработки, объема публикаций российских авторов;
- изменения структуры ВВП и структуры публикаций российских авторов;
- зависимость роста объемов инвестиций в народное хозяйство и рост объемов публикаций (по отраслям народного хозяйства);
- зависимость роста выпуска специалистов государственных и муниципальных вузов и роста объемов публикаций (по отраслям народного хозяйства);
- прогнозировать динамику изменений показателей многомерных, например технико-экономических, объектов и процессов во времени (например, корреляции роста индекса промышленного производства по отношению к предыдущему периоду и прироста инвестиций за тот же период).

Развитие и внедрение методов и средств (продуктов и услуг) постобработки цифровых информационных ресурсов стало бы значительным вкладом как в развитие информатики, так и в становление инновационной экономики в нашей стране, а также в перспективе этот процесс мог бы трансформироваться в новое научное направление – *сетевую мультипликативную аналитической постобработку информации.*

F. ***Создание системы информационной поддержки инновационной деятельности и трансфера технологии.*** Для развития инновационных процессов (в отраслях промышленности) исключительно важна информационная поддержка взаимодействия ключевых аудиторий на этапах трансфера технологий (инновационного цикла «исследование – разработка – производство»). Насущно необходима разработка проблемно-ориентированного интернет-ресурса, обеспечивающего интерактивное взаимодействие и информационную поддержку участников инновационных процессов путем создания единой интегрированной информационной среды отбора, ведения и реализации инновационных проектов. В ВИНИТИ следует разработать интерактивную подсистему, в которую будут включены следующие элементы: *индикативная БД инноваций, БД потенциальных инвесторов (частных и государственных), БД предприятий и организаций*, заинтересованных в поиске и внедрении тех или иных научно-технических разработок.

Концептуальным прототипом такого интернет-ресурса является информационная служба Евросоюза CORDIS – интерактивная информационная платформа в области европейских инноваций, исследований и разработок, которая предоставляет пользователям результаты исследований и разработок по всему инновационному циклу с помощью ряда подсистем, средств и 10 поисковых БД. К настоящему времени в ней зарегистрировано свыше 300 тыс. пользователей.

Сегодня реально функционирует лишь Федеральный портал по научной и инновационной деятельности (www.sci-innov.ru). Его отличительная особенность – ориентация на весьма ограниченную тематику, определяемую перечнем приоритетных направлений развития науки, технологий и техники и перечнем критических технологий РФ. Необходимо отметить, что существенным условием его устойчивого функционирования и развития является организация участия пользователей в информационном наполнении её баз данных. Важно также обеспечение информационного взаимодействия системы информационной поддержки инновационной деятельности и трансфера технологии с государственными информационными системами (ИС РФФИ, ИС РНФ, ИС Федеральной целевой научно-технической программы), хранящими полнотекстовую информацию о результатах исследований (в том числе фундаментальных) и разработок, которые могут иметь дальнейшую промышленную коммерческую реализацию.

G. ***Разработка САПР информационной поддержки работ инновационного цикла.*** В современных условиях для разработки и производства новой продукции актуально и необходимо использование системы автоматизированного проектирования (САПР) информационного обеспечения работ по всему инновационному циклу (так же, как и использование конструкторских САПР, или САПР технологической подготовки производства). Такая система позволит осуществлять проектирование и эффективное управление комплексным информационным обеспечением во взаимосвязи с актуализирующимися задачами и действующими производственными планами по всему распределенному во времени инновационному циклу [13].

Концептуальная основа новой технологии информационного обеспечения работ по всем этапам инновационного цикла должна базироваться на следующих основных положениях:

- автоматизированное проектирование и управление комплексным информационным обеспечением цикла исследование – разработка – производство;
- предоставление комплексной информации, включающей научно-техническую, технико-экономическую, нормативную (в том числе стандарты на изделия и процессы), прогнозно-аналитическую информацию высокой степени обработки и др.;
- функционирование системы в корпоративной интегрированной информационной среде Интранет,

и использование информационных ресурсов сети Интернет;

- количественная оценка уровня организации комплексного информационного обеспечения отдельных этапов и всего инновационного цикла в целом.

Это должно сократить временной лаг и качественно повысить интегральный уровень информационной поддержки процессов создания новой наукоемкой продукции. Опосредовано это будет содействовать повышению качества и конкурентоспособности отечественной высокотехнологичной продукции. Средой функционирования САПР информационной поддержки должна быть специализированная платформа или, что прагматически более предпочтительно, универсальная полнофункциональная автоматизированная информационно-библиотечная система (АИБС) предприятия. На базе АИБС возможно хранение и ведение как внешней технико-экономической и научной информации, так и всех типов информационных ресурсов предприятия, а также удовлетворение различных информационных запросов специалистов, осуществление депозитарной функции и поддержка однородной информационной среды.

H. ***Реализация технологии информационного обслуживания на базе электронного реферативного журнала (с индикативным рефератом) в сети Интернет.*** Задача реферирования мирового потока научной литературы не утратила своего значения, но существенно изменилась. Следует отметить, что в развитых странах по-прежнему издается около 3 тыс. реферативных журналов (РЖ), они выходят в электронном виде и выполняют информационно-поисковые и науковедческие функции, а их рефераты становятся в основном индикативными. Это нейтрализует действие брэдфордовского закона рассеяния публикаций определенной тематики по всему массиву журналов, способствует развитию национальной науки, выработке собственной терминологии и собственной информационной политики.

Основные критические пункты реструктуризации электронного РЖ:

- переориентация на индикативный реферат;
- широкое использование аннотаций (резюме) статей в журнале; возможен минимальный вариант – по каждой статье дается реферат на языке оригинала и русский текст названия и аннотации после машинного перевода (для английского, немецкого, французского языков) с постредактированием;
- автоматическое индексирование статей;
- минимизация временного лага до 1–1,5 месяцев;
- реализация режимов: электронного ИРИ, предоставления данных по произвольным выборкам и срезам, информационного мониторинга (по работам, проектам и/или программам);
- детальная подготовка и проведение, параллельно с традиционной технологией, пилотного цикла с добавлением рисунков, формул, графики в текст реферата.

I. ***Создание доступной через сети общего пользования базы данных по производимой и потребляемой промышленной продукции (ПППП) и стандартам России (стран СНГ, стран БРИКС, ШОС).*** Источниками комплектования БД ПППП будут служить промышленные каталоги и буклеты, материалы выставок, ресурсы Интернета. Эта БД может существенно дополнить информационную поддержку инновационной деятельности. Её прототип – Федеральный фонд промышленных каталогов. Предполагается установить взаимодействие с Министерством промышленности и торговли России, которое работает над созданием Государственной информационной системы промышленности, предусмотренной Федеральным законом от 31.12.2014 № 488-ФЗ (ред. от 13.07.2015) "О промышленной политике в Российской Федерации". «Эта система создается в целях автоматизации процессов сбора, обработки информации, необходимой для обеспечения реализации промышленной политики и осуществления полномочий федеральных органов исполнительной власти по стимулированию деятельности в сфере промышленности, информирования о предоставляемой поддержке субъектам деятельности в сфере промышленности, а также для повышения эффективности обмена информацией о состоянии промышленности и прогнозе ее развития» (Ст.1). Функционирование БД ПППП совместно с Системой поддержки трансфера технологий (п. F) и с БД по кабинету фирм (отечественных и зарубежных) существенно повысило бы уровень информационного обеспечения промышленности РФ.

K. ***Создание государственной вебометрической системы цифрового пространства научных библиотек.*** Современный кризис информационно-библиотечной системы, который приобрел перманентный характер, обусловлен стремительным развитием телекоммуникаций и информационных технологий. В значительной степени Интернет стал важным альтернативным источником общедоступной и специальной информации для ученых, специалистов, населения в целом. Постоянно снижаются показатели основной функциональной деятельности библиотек (как публичных, так и научных): числа читателей, посещаемости, объемов книговыдачи. При этом можно констатировать, что количество обращений к веб-серверам библиотек неуклонно растет в противовес обычным посещениям, частота которых неуклонно снижается. Под воздействием внешних факторов постепенно меняется менталитет потребителя (простого читателя, специалиста, ученого) – ему не нужна книга как таковая, ему нужны *знания*. Для сохранения библиотечной системы как институционального компонента информационной инфраструктуры ГСНТИ вебометрическая система должна стать современным эффективным инструментом развития библиотек в цифровой среде [14].

В качестве основных задач, решаемых в процессе создания вебометрической системы библиотек, выделим:

- повышение роли и значимости публичных и научных библиотек в обществе;
- сохранение и развитие функциональной деятельности библиотек (в зависимости от их типа и вида), поддержание позитивного имиджа в мировом веб-пространстве;

- совершенствование (опосредовано) состава и структуры фондов, оптимизацию комплектования библиотек;
- интенсификацию процессов цифровизации фондов библиотек;
- стимулирование процессов диверсификации библиотечных услуг и продуктов в цифровой среде;
- мониторинг и поддержку принятия управленческих решений;
- социологический мониторинг культурного и образовательного предпочтения россиян;
- формирование интегральной оценки уровня и рейтингового распределения библиотек.

Сбор данных должен осуществляться в автоматическом режиме с использованием API[2] продвинутых поисковых систем (Google, Yahoo, Яндекс).

Web-система должна поддерживать режим постоянного мониторинга web-сайтов библиотек и автоматическую аналитическую постобработку результатов вебометрических исследований с использованием современных методов наукометрии и многомерного анализа данных (для анализа структур научных сайтов следует использовать методы теории графов и метод главных компонентов).

Система должна обеспечивать неэкспертное автоматическое (автоматизированное) формирование сопоставительных, рейтинговых и комплексных оценок, выявление эмпирических закономерностей, получение интегральных характеристик web-сайтов библиотек в режиме квазиреального времени.

L. ***Реализация технологии формирования информационных продуктов прогнозно-аналитического и обзорного характера.*** Важнейшим приоритетом является воссоздание на базе новых информационных технологий традиционного для ВИНИТИ направления переработки информации с выходными продуктами прогнозно-аналитического и обзорного характера. Например, подготовка ежемесячных выпусков предметно-тематических и/или проблемно-ориентированных экспресс-информационных материалов следующей структуры: краткий обзор (10 тыс. знаков); библиографический указатель (75–150 тыс. знаков). Ключевыми задачами являются: определение актуальных тематик и создание условий для привлечения к сотрудничеству квалифицированных специалистов, номинация информационных продуктов и услуг, оценка издержек и расчёт ценообразования.

В целом реализация этих задач помимо приносимых статусных и экономических выгод влияет на расширение возможностей использования результатов прогнозно-аналитической и наукометрической деятельностей в научно-промышленной сфере и управлении народным хозяйством, а также создает реальную основу для:

- сопоставительного анализа структуры и уровня отечественной и мировой науки;
- определения тенденций и процессов в научно-технической сфере;
- прогнозирования развития технологий, экономики, общества;
- выявления точек роста, наиболее актуальных и/или стагнирующих научных направлений;
- мониторинга структуры (программ) отечественного научно-промышленного комплекса.

Перспективной и важной сферой деятельности ВИНИТИ является выполнение целевых заказных работ обзорно-аналитического характера по крупным проектам и комплексным программам.

ОЦЕНКА УРОВНЯ ИНФОРМАЦИОННОГО ОБЕСПЕЧЕНИЯ ЦЕЛЕВЫХ ПРОГРАММ И КРУПНЫХ НАУЧНО-ТЕХНИЧЕСКИХ ПРОЕКТОВ

Актуализация задач развития высокотехнологичных секторов экономики, мировые тенденции сокращения временных характеристик цикла «исследование – разработка – производство», необходимость эффективной реализации национальных проектов РФ на период 2019-2024 гг. – все это выдвигает новые, более высокие требования к организации и качеству информационного обеспечения (ИО). На примере оценки уровня информационного обеспечения целевых комплексных программ (ЦКП), в рамках которых обычно осуществляется создание качественно новых систем и образцов техники, рассмотрим основы методологического подхода к управлению информационным обеспечением. Для этого введем комплексную меру U_{Σ}.

Показатель (коэффициент) охвата комплекса работ научной и технико-экономической информацией по проблемам организаций – участников разработки и реализации одной ЦКП определяется выражением:

$$Q = \sum_{i=1}^{S} \left(g_i \cdot W_{Ni} \cdot W_{Ti}^{-1} \right) , \qquad (1)$$

где: S – количество тематических направлений, по которым необходимо осуществлять информационное обслуживание; W_{Ni} – доля исполнителей комплексной программы, которые получают информацию по *i*-й тематике; W_{Ti} – число исполнителей ЦКП, которым требуется информация по *i*-й тематике; g – приведенный вес *i*-й тематики (по важности, актуальности), $g_i = g_i' \cdot \sum_i g_i \cdot$, где g_i' – вес *i*-й тематики.

Показатель качества информационного обеспечения R определяется как:

$$R = M^{-1} \cdot \sum_{j=1}^{M} \left(K_{0j} \cdot K_j^{-1} \right) \; ; \; M \le S \qquad (2)$$

где: M – количество тематических направлений, по которым ведется информационное обслуживание; K_{0j} – реальный уровень качества информационного обеспечения разработчиков по *j*-му тематическому

[2] API (Application Programming Interface) – это специальный интерфейс или приложения, с помощью которого одна программа/приложение может взаимодействовать с другой; с помощью API различные программы и приложения могут использовать функции и ресурсы друг друга.

направлению; K_j – предельно достижимый (или оптимальный) уровень качества информационного обеспечения разработчиков по *j*-му тематическому направлению.

В общем случае уровень качества определяется с помощью экспертных методов оценок и представляет собой функцию вида:

$$Kj = \varphi(P,O,F,C,D) \ , \qquad (3)$$

где: P – полнота, достоверность и точность представляемой информации; O – оперативность представления вторичных и первичных документов, фактографической информации, данных логической обработки; F – форма представления и выдачи информации; C – уровень логической обработки данных; D – наличие ограничений и доступность информации.

На основании выражений (1) и (3) определяется комплексная мера уровня информационного обеспечения одной ЦКП как:

$$U_1 = Q \cdot R = \\ = M^{-1} \cdot \sum_{j=1}^{M} \left(K_{0j} \cdot K_j^{-1} \right) \cdot \sum_{i=1}^{S} \left(g_i \cdot W_{Ni} \cdot W_{Ti}^{-1} \right) \le \alpha_0 \ , \qquad (4)$$

где: α_0 – заданный или оптимальный, или предельно достижимый уровень информационного обеспечения.

С помощью выражений 1-4 комплексная мера U_Σ уровня информационного обеспечения совокупности целевых программ (например одной отрасли) определяется как:

$$U_\Sigma = L^{-1} \cdot \sum_{i=1}^{L} \left(Q_l \cdot R_l \right) \ , \qquad (5)$$

Используя выражения (4) и (5), можно дать как абсолютную, так и относительную оценку уровня информационного обеспечения отдельных ЦКП (крупных проектов и/или инновационных этапов) и их совокупности, например, по отношению к тому же уровню в других ЦКП или отраслях.

Разработка методологии комплексной оценки уровня информационного обеспечения требует итерационного решения целого ряда задач связанных, прежде всего, с развитием экспертных методов оценок, формализации и определения численных значений и единиц измерения разнородных компонент (P,O,F,C,D) и их корректной нивелировки, а также накопления статистических данных для проведения сопоставительного анализа.

ЗАКЛЮЧЕНИЕ

В современных экономических условиях проблема преодоления тенденций инерционного развития национальной информационной системы требует для своего решения новых идей, новых концептуальных подходов, новых технологий, новых информационных продуктов и услуг. Особую значимость и актуальность приобретает производство информационных продуктов и услуг на основе аналитической постобработки информации, технико-экономического анализа объектов и процессов в различных разрезах, многоаспектного моделирования и прогнозирования. Широкое применение в структуре ГСНТИ цифровых информационных ресурсов, новых информационных технологий содействует более эффективному решению задач информационного обеспечения фундаментальных и прикладных исследований, инновационной деятельности [15], что обеспечивается рациональным, сбалансированным развитием информационной инфраструктуры, информационных ресурсов, информационных технологий. Комплекс задач по более глубокой переработке информации, извлечения новых знаний является частью общих проблем информационного обеспечения научных исследований и разработок и имеет определенное экономическое измерение [16]. Информационный компонент научно-технического комплекса России прямо или косвенно отражается в эффекте:

- мультипликации использования новых научно-технических результатов, знаний и информационных ресурсов;
- комплексного подхода к инвестициям и инновациям в научно-промышленной сфере;
- экономии общественно необходимого времени и материально-технических ресурсов за счет типовых проектных решений;
- трансфера технологий и использования частных технических решений (в разных отраслях).

Для реализации масштабных задач модернизации национальной информационной инфраструктуры необходимо привлечение дополнительных средств из госбюджета, государственных и целевых федеральных программ, научных и технологических фондов. В частности, такие перспективные направления как технологии Большие Данные (Big Data) и широкополосный Интернет в настоящее время существенно отстают от мирового уровня. Представляется целесообразным сформировать для ВИНИТИ РАН, как головной организации ГСНТИ, целевое госзадание на общую координацию работ, разработку «дорожной карты», развитие общесистемной нормативно-методической базы, мониторинг формирования и использования цифровых информационных ресурсов, проведение научных исследований проблем сбора, аналитико-синтетической переработки, хранения, поиска, распространения и использования научно-технической информации. Следует отметить, что вне рамок настоящей работы (в силу статейных ограничений) остались вопросы, актуальные для современной российской экономики, связанные с задачами информационного обеспечения процессов импортозамещения в сфере высоких технологий, а также с необходимостью частичной конверсии оборонной промышленности для производства продукции гражданского назначения. В заключение следует отметить, что в России имеется необходимый научный, технический и экономический потенциал для формирования соответствующей требованиям времени информационной инфраструктуры, и этот потенциал необходимо реализовать в кратчайшие сроки.

СПИСОК ЛИТЕРАТУРЫ

1. Антошкова О.А., Белоозеров В.Н., Дмитриева Е.Ю., Шапкин А.В. Разработка онтологии НТИ на основе библиографических классификаций // Информационное обеспечение науки: новые технологии: Сб. научн. трудов / ред. Н.Е. Каленов, В.А. Цветкова. – Москва : БЕН РАН, 2017. – С. 292-300. – ISBN 978-5-201-13141-8.
2. Антошкова О.А., Белоозеров В.Н., Дмитриева Е.Ю. и др. Построение онтологии информационных ресурсов в виде сети библиографических классификаций // Материалы научно-практической конференции «Перспективные направления научных исследований и критические технологии в классификационных системах» (25-27 октября 2017 г.). – Москва: ВИНИТИ РАН, 2017. – 96 с. – С. 20-25. – ISBN 978-5-94577-072-0.
3. Сюнтюренко О.В., Белоозеров В.Н., Дмитриева Е.Ю. и др. Сеть классификаций по науке и технике как механизм смысловой навигации и поиска информации в пространстве знаний // Депонировано в ВИНИТИ РАН 19.12.2019, № 120-В2019.
4. Сюнтюренко О.В. Перспективы использования интернет-СМИ, журналов открытого доступа и социальных медиа // Научно-техническая информация. Сер. 1. – 2015. – № 6. – С. 30-36; Syuntyurenko O.V. Prospects for Using Online Media, Open-Access Journals, and Social Media Networks in the Field of Science and Technology // Scientific and Technical Information Processing. – 2015. – Vol. 42, № 2. – P. 112-118.
5. Биктимиров М.Р., Гиляревский Р.С., Сюнтюренко О.В. Новая концептуальная основа развития информационной деятельности ВИНИТИ РАН // Научно-техническая информация. Сер. 1. – 2016. – № 1. – С. 1-8; Biktimirov M.R., Gilyarevskii R.S., Syuntyurenko O.V. A New Conceptual Basis for the Development of the Information Activities of the All-Russian Institute for Scientific and Technical Information of the Russian Academy of Sciences // Scientific and Technical Information Processing. – 2016. – Vol. 43, № 1. – P. 1-7.
6. Brynjolfsson E., McAfee A. The Second Machine Age: Work, Progress, and Prosperity in a Time of Brilliant Technologies. – New York: Norton & Company, 2016. – 320 p.
7. Дроздова К.А. Машинный перевод: история, классификация, методы // Материалы III междунар. науч. конф. «Филологические науки в России и за рубежом» (Санкт-Петербург, июль 2015 г.). – СПб: Свое издательство, 2015. – С. 139-141. – URL https://moluch.ru/conf/phil/archive/138/8497/ (дата обращения: 28.12.2018).
8. Колганов Д.С., Данилов Е.А. Обзор аналитической, статистической и нейронной технологии машинного перевода // Международный студенческий научный вестник. – 2018. – № 3-2. – URL: http://eduherald.ru/ru/article/view?id=18262 (дата обращения: 28.12.2018).
9. Борисова Л.Ф., Сюнтюренко О.В. Реферативный банк данных ВИНИТИ РАН: перспективы постобработки информации с использованием методов анализа данных // Научно-техническая информация. Сер. 1. – 2007. – № 11. – С. 6-11.
10. Калачихин П.А. Принципы построения государственной наукометрической системы // Научно-техническая информация. Сер. 2. – 2016. – № 7. – С. 11-23; Kalachikhin P.A. The Principles of the Design of the State Scientometric System // Automatic Documentation and Mathematical Linguistics. – 2016. – Vol. 50, № 4. – P. 161-172.
11. Сюнтюренко О.В., Гиляревский Р.С. Использование методов наукометрии и сопоставительного анализа данных для управления научными исследованиями по тематическим направлениям // Научно-техническая информация. Сер. 2. – 2016. – № 12. – С. 1-12.
12. Сюнтюренко О.В. Цифровая среда: аналитическая постобработка информации с использованием методов наукометрии и анализа данных // Научно-техническая информация. Сер. 1. – 2019. – № 4. – С. 8-16; Syuntyurenko O.V. Digital Environment: Information Analytical Postprocessing Using the Scientometric and Data Analysis Methods // Scientific and Technical Information Processing. – 2016. – Vol. 46, № 2. – P. 59-66.
13. Сюнтюренко О.В., Булычева О.С. Концептуальный облик перспективного технологического пакета информационной поддержки наукоемкого производства // Научно-техническая информация. Сер. 2. – 2016. – № 4. – С. 1-10; Syuntyurenko O.V., Bulycheva O.S. The Conceptual Form of an Advanced Technology Package for the Information Support of Knowledge-Intensive Production // Automatic Documentation and Mathematical Linguistics. – 2016. – Vol. 50, № 2. – P. 47-55.
14. Булычева О.С., Сюнтюренко О.В. Концептуальные положения и предпосылки создания вебометрической системы цифрового пространства библиотек // Сборник Президентской библиотеки. Сер. «Электронная библиотека». – 2018. – Вып. 8. – С. 19-31.
15. Сюнтюренко О.В., Дмитриева Е.Ю. Государственная система научно-технической информации в структуре задач цифровой экономики // Научно-техническая информация. Сер. 1. – 2019. – № 9. – С. 1-11. Syuntyurenko O.V. Dmitrieva E.Yu. The State System for Scientific and Technical Information within the Objectives of the Digital Economy // Scientific and Technical Information Processing. – 2016. – Vol. 46, № 4. – P. 288-297.
16. Родионов И.И., Гиляревский Р.С., Цветкова В.А. Информационная деятельность как инфраструктура национальной экономики. – СПб: Издательство, 2016. – 200 с.

Материал поступил в редакцию 14.07.20.

Сведения об авторах

ДМИТРИЕВА Елена Юрьевна – кандидат технических наук, заведующая научно-методологическим отделением ВИНИТИ РАН,
e-mail: niipio@mail.ru

СЮНТЮРЕНКО Олег Васильевич – доктор технических наук, профессор, ведущий научный сотрудник ВИНИТИ РАН, БЕН РАН,
e-mail: olegasu@mail.ru

ИНФОРМАЦИОННЫЙ АНАЛИЗ

УДК 050: [002:001.8]

Р.С. Гиляревский, А.Н. Либкинд, В.Г. Богоров, И.А. Либкинд

Вычисление периода полужизни научных журналов в условиях неполноты данных *Journal Citation Reports**

Решается задача определения динамики показателей периода полужизни научных журналов Cited Half-life (CdHL) и Citing Half-life (CgHL)) в условиях существенной неполноты и недостаточной точности данных, содержащихся в Journal Citation Reports (JCR). Выполнен детальный анализ этих данных для показателей CdHL и CgHL за период 1997–2018 гг. Уточнена ранее сформулированная гипотеза о подобии распределений журналов по показателям периода полужизни. Проработаны методы определения средневзвешенных значений соответствующих показателей. Проверка гипотезы подобия позволила определить временные рамки, в пределах которых ее справедливость не вызывает сомнений. Показано, что динамика значений показателей периода полужизни положительна, причем эта динамика значительно более ярко выражена для постоянно сохраняющихся журналов, т.е. таких журналов, каждый из которых присутствовал в JCR в течение всего 22-х летнего исследуемого периода.

Ключевые слова: *старение литературы, период полужизни, Cited Half-life, Citing Half-life, Journal Citation Report, распределения журналов, гипотеза подобия*

DOI: 10.36535/0548-0027-2020-11-2

ВВЕДЕНИЕ

Когда Дж. Бернал в 1958 г. высказался о желательности измерения скорости старения литературы, а Р. Бартон и Р. Кеблер в 1960 г. использовали для этого показатель периода полужизни статей, они, в первую очередь, интересовались изучением темпов развития различных областей науки. В то время для этого приходилось затрачивать много сил, поскольку источники ссылок на статьи были традиционными, и прослеживать их приходилось вручную. Данные, опубликованные в [1], до сих пор приводятся в качестве примера старения литературы в различных науках, хотя они давно уже устарели. Теперь мы располагаем оцифрованными источниками о цитировании научных статей, среди которых наиболее надежным и авторитетным служит *Journal Citation Reports (JCR)*.

Однако этот источник долгое время был ограничен ретроспективой старения в 10 лет, после чего отсутствие точных данных обозначалось текстовым выражением вида «>10». Это легко понять, если предположить, что число номеров каждого из 10 тыс. журналов, число статей в каждом номере, число ссылок в каждой статье увеличивается в 10 раз (хотя на самом деле в *JCR* эта величина выше). Тогда число ссылок, которое надо проследить за 10 лет, составит 100 млн. Правда, с 2017 г. *JCR* это ограничение снял, что позволило разработать в рамках нашего проекта методику подсчета, «восстанавливающую» полную ретроспективу значений показателей периода полужизни [2]. Тем не менее, если учитывать конечную цель таких подсчетов в вычислении периода полужизни статей в отраслях знания, то можно сказать, что трудности на этом не заканчиваются. *JCR* публикуется ежегодно в двух различных по тематике выпусках, в которых журналы частично дублируются, поскольку статьи в них могут относиться к разным тематическим категориям. Да и сами эти тематические категории в Web of Science (WoS) и JCR с течением времени изменяются, а также могут переходить из одного тематического выпуска в другой. Некоторые журналы в определенные годы могут исчезать из *JCR* из-за временного снижения импакт-фактора или по другим причинам.

Настоящая работа посвящена преодолению названных трудностей для полноценного исследования

* Работа выполнена во исполнение государственного задания ВИНИТИ РАН по теме 0003-2019-0001 и при поддержке Российского фонда фундаментальных исследований (проекты РФФИ 20-07-00014 и 20-010-00179.

динамики показателей полужизни журнальных статей – *Cited Half-life* (*CdHL*) и *Citing Half-life* (*CgHL*). Показатель *CdHL* (период полужизни цитируемых статей журнала) определяется на основе данных о годе опубликования тех статей[1] из определенного журнала, которые в заданные годы[2] были процитированы другими периодическими изданиями. Соответственно, показатель *CgHL* (период полужизни цитирующих статей журнала) вычисляется на основе данных о годе опубликования тех статей из других журналов, которые в заданные годы[3] были процитированы этим журналом. Полученные сведения помогут проследить тенденции развития отраслей знания для последующего прогнозирования перспективности тематики экспериментальных исследований.

Сама идея определения динамики показателей периода полужизни основывается на том, что для соответствующего набора журналов (например, категорий JCR и/или их совокупностей) того или иного ежегодного выпуска *JCR* необходимо вычислить некоторое обобщающее для этого распределения значение показателя. В этом качестве естественно использовать его средневзвешенное значение, которое учитывает не только все показатели в распределении, но и вес (число журналов, их долю) каждого из них. Расположив значения по годам (в таблице или на графике), мы сможем увидеть динамику соответствующего показателя.

ИСХОДНЫЕ ДАННЫЕ, ИХ ОСОБЕННОСТИ И ПРЕДВАРИТЕЛЬНАЯ ОБРАБОТКА

В качестве источников исходных данных были использованы ежегодные выпуски аналитико-статистической базы данных *Journal Citation Reports* (*JCR*) за период 1997–2018 гг. Эта БД расположена на платформе *Web of Science* (*WoS*) компании *Clarivate Analytics*. *JCR* публикуется ежегодно в виде двух тематических выпусков: *Journal Citation Reports – Science Edition* (*JCR-SE*) и *Journal Citation Reports – Social Science Edition* (*JCR-SSE*). В основном *JCR* представляет результат статистической обработки двух журнальных баз данных *WoS*. А именно: *JCR-SE* – результат обработки БД *Citation Index-Expanded – SCI-E* (естественные, точные и технические науки) и *JCR-SSE* – результат обработки БД *Social Sciences Citation Index – SSCI* (общественные науки)[4]. Каждый из тематических выпусков содержит характеристики соответствующих единиц двух важнейших информационных объектов (на языке теории баз данных – сущностей):

1) информационный объект «Журнал»: в качестве единицы (экземпляра) этого информационного объекта выступает один из журналов, включенных в *JCR*;

2) информационный объект «Тематическая категория *WoS*»: в качестве единицы (экземпляра) этого информационного объекта выступает одна из категорий *WoS*, включенных в *JCR*.

Прежде чем перейти к детальному описанию каждого из рассматриваемых информационных объектов, приведем ситуации, которые были обнаружены при анализе значений показателей периода полужизни указанных информационных объектов и их единиц. Обнаруженные ситуации можно сгруппировать следующим образом:

a) *«раздвоение» отдельных единиц объектов*. Раздвоение 1-го типа – это ситуация, когда в одном и том же году журнал или категория *WoS* оказываются представленными одновременно и в *JCR-SE*, и в *JCR-SSE в одной и той же категории WoS*. Раздвоение 2-го типа – это ситуация, когда в одном и том же году журнал оказывается представленным одновременно в нескольких категориях *WoS;*

b) *миграция отдельных единиц объектов.* Это понятие целесообразно применять только для описания «поведения» категорий *WoS*: ситуация, когда некоторая категория *WoS*, находясь в определенном году в конкретном тематическом выпуске, через несколько лет перемещается в другой тематический выпуск, например из *JCR-SE* в *JCR-SSE* или наоборот;

---

[1] Если придерживаться терминологии *JCR*, то следует сказать, что помимо статей учитываются и некоторые другие публикации, а именно такие, которые в принципе могут быть процитированными (*citable items*), в частности, обзоры.

[2] Обычно это годы, следующие за годом опубликования первого ежегодного комплекта цитирующих журналов.

[3] Обычно это годы, следующие за годом опубликования первого ежегодного комплекта цитируемых журналов.

[4] Именно эти две БД, согласно *Clarivate Analytics*, являются единственными источниками исходных данных для *JCR*. Так, на сайте *Clarivate Analytics* в разделе «*Journal Citation Reports Help*» читаем: «*Journal Citation Reports aggregates the meaningful connections of citations created by the research community through the delivery of a rich array of publisher-independent data, metrics and analysis of the world's most impactful journals included in the Science Citation Index Expanded* (*SCIE*) *and Social Sciences Citation Index* (*SSCI*) …» (http://jcr.help.clarivate.com/Content/home.htm).

Однако, как показывает анализ, значительная часть журналов, отражающихся в БД *Arts & Humanities Citation Index–A&HCI* (искусство и гуманитарные науки) также представлены в *JCR*. Так, например выпуск *JCR* за 2018 г. содержит 95 журналов, соответствующих категории «*History*» (назовем эти журналы «журналами по истории»), которая согласно классификации *Web of Science* (*WoS*), включена в БД *A&HCI*. Причем более половины этих журналов (53 из 95) *WoS* относит к единственной категории, а именно, к категории «*History*» и только к этой категории. Из оставшихся 42 журналов по истории четыре относятся также к категории «*Cultural Studies*», и два – к категории «*Ethic Studies*» (обе категории соответствуют гуманитарной проблематике). Таким образом, из 95 журналов по истории более 59 (53+6), т.е. 62% представлены в *WoS* только в БД *A&HCI*. Добавим, что из 4739 наименований журналов, которые были включены хотя бы в один ежегодный выпуск *JCR-SSE* за период 1997-2018 гг., 264 журнала согласно классификации *WoS* были отнесены только к гуманитарным наукам, а общее число журналов, которые посвящены, прежде всего, гуманитарным проблемам, за это период составило 315 наименований. Следовательно, используя *JCR*, мы можем быть уверены, что в анализ попадают и чисто гуманитарные тематики и соответствующие им журналы.

ЗАМЕЧАНИЕ. Возможно, что мнение о том, что *JCR* не содержит журналы из БД *A&HCI* связано с тем, что в *JCR* для этих журналов не приводятся значения импакт-фактора. В связи с этим отметим, что в *JCR* значения остальных показателей, которые в последние годы являются «стандартными» для этого издания (и что особенно важно для целей настоящей статьи – показатели периода полужизни), как правило, приводятся.

c) *полное отсутствие данных*. Ситуация, когда в соответствующих полях *JCR* вместо числового значения показателя указывается «*Not Available*» («Данные недоступны» или «Данные отсутствуют»);

d) *неточные (недостаточно точные) данные*. Ситуация, когда в соответствующих полях *JCR* для периода полужизни (в случае, если значение показателя превышает 10 лет) вместо конкретного числового значения приводится текстовое выражение вида «>10.0».

Журналы

Число журналов в ежегодном выпуске *JCR-SE* за рассматриваемый период почти удвоилось: 4962 названий в 1997 г. и 9156 – в 2018 г. Аналогичная тенденция характерна и для *JCR-SSE*: 1672 и 3382 журнала в 1997 г. и в 2018 г. соответственно.

Журнал в годовом выпуске *JCR* может соответствовать одной или более категориям *WoS*, т.е. между журналом и категориями *WoS* в *JCR* установлено отношение вида «один-ко-многим» (раздвоение 2-го типа). Естественно, что журнал может присутствовать также и в различных тематических ежегодных выпусках *JCR* (раздвоение 1-го типа): как в *JCR-SE*, так и в *JCR-SSE*. Очевидно, что эти ситуации, т.е. раздвоение 1-го и 2-го типов, для журналов совершенно естественны и являются проявлением многоаспектности исследований, публикуемых в таких журналах. Отметим, что это не препятствует дальнейшей обработке данных.

Таблица 1

***JCR-SE*: характеристики массивов журналов с точки зрения возможностей оценки периода полужизни изданий по естественным, точным и техническим наукам**

Годы *JCR*	Все журналы	*Для Cited Half-life (CdHL)*					*Для Citing Half-life (CgHL)*				
		Все журналы, %		Абсолютно сохранившиеся журналы, %			Все журналы, %		Абсолютно сохранившиеся журналы, %		
1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
1997	4962	**15,7**	13,6	68,8	14,5	9,6	**30,0**	4,8	68,8	29,9	0,8
1998	5467	**15,3**	16,0	62,4	15,6	7,6	**29,3**	4,0	62,4	30,2	0,5
1999	5550	**15,2**	13,9	61,5	16,1	5,3	**29,6**	4,3	61,5	30,5	0,6
2000	5686	**15,1**	13,3	60,0	16,7	4,2	**30,6**	4,4	60,0	32,6	0,6
2001	5752	**15,1**	12,1	59,3	17,0	3,3	**30,9**	3,9	59,3	32,7	0,4
2002	5876	**15,5**	10,6	58,1	18,1	2,5	**31,5**	3,7	58,1	33,5	0,4
2003	5907	**15,9**	9,3	57,8	19,0	1,9	**31,9**	3,4	57,8	34,4	0,4
2004	5969	**16,3**	8,2	57,2	19,5	1,8	**31,2**	3,1	57,2	33,5	0,6
2005	6088	**16,3**	7,3	56,0	20,2	1,5	**31,7**	2,5	56,0	34,7	0,5
2006	6166	**16,5**	6,4	55,3	20,8	1,2	**31,6**	2,3	55,3	35,2	0,5
2007	6426	**16,1**	6,4	53,1	21,2	0,9	**31,3**	4,2	53,1	34,9	2,4
2008	6620	**16,6**	5,5	51,5	22,6	0,7	**31,5**	1,8	51,5	36,5	0,4
2009	7387	**16,3**	8,6	46,2	24,2	0,4	**31,8**	1,9	46,2	37,0	0,4
2010	8073	**15,6**	9,7	42,3	24,8	0,4	**32,7**	1,9	42,3	38,0	0,3
2011	8336	**16,1**	8,5	40,9	27,1	0,3	**32,9**	2,1	40,9	39,6	0,5
2012	8471	**16,5**	7,1	40,3	28,9	0,2	**34,1**	1,7	40,3	42,2	0,3
2013	8539	**17,4**	5,4	40,0	30,9	0,2	**36,2**	1,9	40,0	44,8	0,4
2014	8659	**18,6**	4,3	39,4	33,6	0,1	**29,6**	1,7	39,4	32,4	0,7
2015	8802	**19,5**	3,5	38,8	36,3	0,2	**30,8**	1,6	38,8	33,9	0,6
2016	8866	**21,4**	**2,3**	38,5	39,4	0,1	**30,9**	1,7	38,5	**34,5**	0,8
2017	*9015*	***22,5***	*1,9*	*37,8*	*42,4*	*0,1*	***30,1***	*1,4*	*37,8*	*34,6*	*0,9*
2018	*9156*	***23,3***	*1,4*	*37,3*	*44,3*	*0,1*	***29,7***	*1,2*	*37,3*	*34,3*	*0,9*

ПРИМЕЧАНИЕ: **1** – годы *JCR*; **2** – общее число журналов; **3** и **8** – неточные данные (общий массив журналов): доля журналов со значениями показателя полужизни, превышающими 10 лет (от общего числа журналов); **4** и **9** – полное отсутствие данных (общий массив журналов): доля журналов, данные о значениях показателя полужизни которых отсутствуют (от общего числа журналов); **5** и **10** – доля абсолютно сохранившихся журналов (от общего числа журналов); **6** и **11** – неточные данные (абсолютно сохранившееся журналы): доля журналов со значениями показателя полужизни, превышающими 10 лет (от числа абсолютно сохранившихся журналов); **7** и **12** – полное отсутствие данных (абсолютно сохранившиеся журналы): доля журналов, данные о значениях показателя полужизни которых отсутствуют (от числа абсолютно сохранившихся журналов).

Таблица 2

***JCR-SSE* : характеристики массивов журналов с точки зрения возможностей оценки периода полужизни изданий по общественным наукам**

Годы *JCR*	Общее число журналов	*Для Cited Half-life (CdHL)*					*Для Citing Half-life (CgHL)*				
		Все журналы, %	Абсолютно сохранившиеся журналы, %				Все журналы, %		Абсолютно сохранившиеся журналы, %		
1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
1997	1672	**12,1**	32,6	73,0	13,1	25,6	**25,3**	5,3	73,0	26,0	1,2
1998	1678	**12,3**	29,3	72,8	13,5	21,9	**26,6**	3,4	72,8	27,4	0,8
1999	1699	**13,9**	26,3	71,9	15,6	18,4	**26,3**	3,9	71,9	27,0	0,9
2000	1697	**15,0**	23,4	72,0	16,7	15,2	**30,2**	4,0	72,0	31,4	1,2
2001	1682	**16,9**	21,5	72,6	19,0	13,8	**32,4**	4,4	72,6	34,0	1,5
2002	1709	**17,9**	19,0	71,4	20,5	11,1	**33,5**	3,2	71,4	36,4	0,6
2003	1714	**19,5**	17,2	71,2	22,5	9,3	**35,5**	3,2	71,2	37,8	0,7
2004	1712	**20,6**	14,6	71,3	23,8	7,9	**37,1**	2,5	71,3	39,4	0,8
2005	1747	**21,3**	13,2	69,9	25,1	7,4	**38,9**	1,8	69,9	42,3	0,7
2006	1768	**22,1**	11,0	69,1	26,6	6,1	**39,9**	1,5	69,1	44,2	0,4
2007	1866	**22,0**	11,7	65,4	28,2	5,2	**40,1**	1,3	65,4	45,4	0,3
2008	1985	**22,9**	10,7	61,5	30,3	3,8	**41,6**	1,5	61,5	47,7	0,3
2009	2257	**25,1**	14,7	54,1	37,1	2,5	**43,3**	1,4	54,1	51,4	0,4
2010	2731	**20,3**	18,6	44,7	34,6	2,6	**41,3**	2,3	44,7	51,1	0,5
2011	2966	**20,4**	18,7	41,2	36,9	2,5	**44,1**	2,4	41,2	55,3	0,5
2012	3047	**21,3**	16,5	40,1	39,1	2,3	**45,8**	2,4	40,1	57,7	0,7
2013	3080	**23,3**	14,8	39,6	42,8	2,5	**51,6**	2,1	39,6	61,7	1,1
2014	3154	**24,8**	12,8	38,7	45,9	2,0	**46,9**	1,5	38,7	49,2	0,6
2015	3224	**27,0**	10,3	37,9	49,2	1,8	**48,1**	1,6	37,9	52,2	0,7
2016	3241	**31,3**	7,6	37,7	54,1	1,1	**50,7**	1,9	37,7	54,3	0,8
2017	*3312*	***33,9***	*5,3*	*36,9*	*59,1*	*0,8*	***46,8***	*1,1*	*36,9*	*49,4*	*0,7*
2018	*3382*	*35,3*	*4,1*	*36,1*	*60,8*	*0,9*	***48,4***	*1,2*	*36,1*	*52,7*	*1,0*

ПРИМЕЧАНИЕ: **1** – годы *JCR*; **2** – общее число журналов; **3** и **8** – неточные данные (общий массив журналов): доля журналов со значениями показателя полужизни, превышающими 10 лет (от общего числа журналов); **4** и **9** – полное отсутствие данных (общий массив журналов): доля журналов, данные о значениях показателя полужизни которых отсутствуют (от общего числа журналов); **5** и **10** – доля абсолютно сохранившихся журналов (от общего числа журналов); **6** и **11**– неточные данные (абсолютно сохранившееся журналы): доля журналов со значениями показателя полужизни, превышающими 10 лет (от числа абсолютно сохранившихся журналов); **7** и **12** – полное отсутствие данных (абсолютно сохранившиеся журналы): доля журналов, данные о значениях показателя полужизни которых отсутствуют (от числа абсолютно сохранившихся журналов)

Полное отсутствие данных в случае журналов. Сведения об этой ситуации, полученные нами в результате соответствующего анализа, приводятся для *JCR-SE* в табл. 1 (графы 4 и 9), а для *JCR-SSE* – в табл. 2 (графы 4 и 9). Для показателя *CdHL* доля таких журналов в *JCR-SE* достигает 13,6%. Для выпусков *JCR-SSE* доля журналов с полным отсутствием данных для показателя *CdHL* достигает почти трети (32,6%). Следует отметить, что с течением времени доля таких журналов заметно уменьшается. Так, если в случае *JCR-SE* для *CdHL* в выпуске 1997 г. эта доля составляет 13,6%, в 2010 г. – 9,7%, а в выпуске 2018 г. – 1,4%. В случае *JCR-SSE* в выпуске 1997 г. – 32,6%, в 2010 г. – 18,6%, а в выпуске 2018 г. – 4% (137 журнала из 3382). Аналогичная картина наблюдается для обоих тематических выпусков *JCR* и в случае другого показателя периода полужизни – *CgHL*, хотя цифры здесь значительно меньше. Так, максимальное значение доли журналов, у которых отсутствовали данные для показателя *CgHL* для случая *JCR-SE* – 4,8% (1997 г.), минимальное – 1,4%. Аналогичные цифры для *JCR-SSE* близки: 5,3% и 1,2% соответственно. Такое падение совершенно естественно: *JCR* со временем удается собрать информацию о недостающих данных.

Неточные данные в случае журналов. Во всех ежегодных выпусках за период 1997–2016 гг. при значениях того или иного показателя, превышающих

10 лет, в соответствующем поле просто указывается текстовое выражение вида «>10». Таким образом, только данные в двух ежегодных выпусках (2017 г. и 2018 г.) из 22 (1997–2018 гг.), попавших в анализ тематических выпусков, не страдают этим недостатком. Доля журналов с неточными данными в любом ежегодном выпуске всегда больше 10%, нередко составляет десятки процентов, в отдельных случаях превышает 50% и, что очень важно, со временем, как правило, увеличивается (см. графы 3 и 8 в табл. 1 и 2). Следует отметить, что тогда, когда эти показатели искусственно не ограничены числом «10», они нередко очень значительно превышают указанное ограничение в 10 лет. Так, для выпуска *JCR-SE* за 2017 г. максимальное значение *CdHL* составляет 105,1 года (!), а для *CgHL* – 113,2 (!!) года. В случае *JCR-SSE* эти цифры ниже, но также достаточно значительны: для *CdHL* – 56,9 года и для *CgHL* – 74 года.

Для настоящего исследования представляет интерес и такой признак журнала, как страна его издания. Как показал анализ, существуют ситуации, когда в разные годы для одного и того же журнала в *JCR* в поле «*Region*», указаны разные страны. Это, скорее всего, объясняется следующим. Данные о журнале и, следовательно, о стране его издания, поступают в *JCR* из заявки издателя журнала в компанию *Clarivate Analytics*. У журналов, особенно международных, иногда происходит смена издателя. Новый издатель может располагаться в стране, отличной от страны нахождения прежнего издателя. В итоге в заявке нового издателя журнала в *Clarivate Analytics* и, следовательно, в описании журнала в *JCR* появится страна, отличная от страны, указанной в предыдущих ежегодных выпусках. С тем, чтобы избежать ненужной неопределенности, в качестве признака «Страна» того или иного журнала мы приняли актуальное значение этого признака, т.е. то, которое приводится в *JCR* за 2018 г.

Тематические категории *WoS*

Ежегодные выпуски *JCR-SE* включают 172–178 тематических категорий *WoS*, а выпуски *JCR-SSE* – 54–58 категорий *WoS*[5] (см. табл. 3)**.** При этом набор категорий в ежегодных выпусках *JCR* со временем меняется, однако эти изменения незначительны.

Раздвоение категорий. В случае категорий может иметь место раздвоение 1-го типа, т.е. ситуация, когда одна и та же категория в одном и том же ежегодном выпуске находится и в *JCR-SE*, и в *JCR-SSE*. Например, в 2018 г. шесть категорий – «*Nursing*», «Psychiatry», «*History & Philosophy of Science*», «*Public, Environmental & Occupational Health*», «*Rehabilitation*», и «*Substance Abuse*» – присутствуют и в *JCR-SE*, и в *JCR-SSE*. Отметим, что с точки зрения логической структуры науки эта ситуация может быть вполне естественной. Однако при обработке таких данных возникает ряд трудностей, и мы, исходя из тематической классификации, принятой в *Web of Science (классификация WoS)*, «привязали» каждую такую категорию к тому тематическому выпуску *JCR*, которому согласно указанной классификации соответствует рассматриваемая категория: к *JCR-SE* или к *JCR-SSE*. В пользу такой привязки свидетельствует, например, следующий факт: подавляющее большинство журналов «раздваивающихся» категорий присутствует в одном и том же году и в *JCR-SE*, и в *JCR-SSE*. Например, в *JCR-SSE* категории «*History & Philosophy of Science*» (*H&PS*) в 2018 г. соответствует 46 журналов, в *JCR-SE* в этом же году – 62 журнала. При этом 37 журналов оказались общими. Таким образом, общее количество уникальных (неповторяющихся) журналов в «объединенной», теперь привязанной только к тематическому выпуску *JCR-SSE* категории *H&PS*, увеличится и составит 71 наименование (46 + (62–37)).

Миграция категорий WoS. Анализ выпусков *JCR* показал также, что встречаются случаи, когда категория *WoS*, находясь в одном году в определенном тематическом выпуске, через несколько лет его «покидает» и «мигрирует» в другой. Миграция категорий, в отличие от их раздвоения, только с большой натяжкой может быть объяснена логикой развития науки. Скорее всего, это результат определенной непоследовательности решений менеджмента *JCR*. Как и в случае с раздвоением категорий, такая мигрирующая категория, исходя из тематической классификации, принятой в *WoS*, была нами «привязана» к соответствующему тематическому выпуску *JCR*.

Полное отсутствие данных о показателях для категорий. В период 1997 – 2002 гг. у всех категорий *WoS* данные о значениях показателей периода полужизни просто отсутствуют: в соответствующих полях для каждой категории *WoS* в *JCR* указано «*Not Available*». Однако, начиная с 2003 г. и в *JCR-SE*, и в *JCR-SSE* ни для одной из категорий таких записей нет.

Неполные (неточные, приближенные) значения данных о показателях для категорий. В период 2003–2018 гг. для всех категорий, у которых значения *CdHL/CgHL* превышают 10 лет, вместо точного числового значения указывается текстовое выражение вида «>10» (напомним, что для периода 1997–2002 гг. эти данные полностью отсутствуют). Доля категорий с такими неточными данными со временем возрастает. Так, в *JCR-SE* для показателя *CdHL* эта доля в 2003 г. составляла 5,9 %, в 2018 г. – 13,5%., а для показателя *CgHL* – 10,6% и 16,3% соответственно. Для *JCR-SSE* доли категорий, которые характеризуются указанной неточностью значений показателей, оказались еще значительнее, а темпы возрастания этой доли выше, чем для *JCR-SE*: в 2003 г. в *JCR-SSE* эти цифры составляла 18,7 %, а в 2018 г. – 42,9%, а для показателя *CgHL* – 14,8% и 42,9% соответственно (см. табл. 3).

Заканчивая рассмотрение данных, которые в *JCR* приводятся для категорий *WoS*, нужно отметить, что самый существенный их недостаток – это то, что каждый ежегодный выпуск обязательно содержит категории, у которых вместо значений *CdHL/CgHL* указано «>10». Причем со временем число и доля таких категорий увеличивается. Эти обстоятельства, к сожалению, делают совершенно невозможным использование таких данных при расчете средневзвешенных показателей *CdHL/CgHL* с помощью описанного далее «Метода виртуализации распределений».

[5] 80–85% категорий *JCR-SSE* приходится на общественные науки и 15–20% – на гуманитарные науки

Таблица 3

Количество категорий *WoS*, включенных в *JCR* и их доля с неточными данными

Год *JCR*	*Journal Citation Report – Science Edition (JCR-SE)*			*Journal Citation Report – Social Science Edition (JCR-SE)*		
	Всего категорий *WoS*	Доля категорий с неточными данными (текстовые значения «>10.0»), %		Всего категорий *WoS*	Доля категорий с неточными данными (текстовые значения >10), %	
		CdHL	*CgHL*		*CdHL*	*CgHL*
2003	170	5,9	10,6	54	16,7	14,8
2004	170	4,7	10,0	54	18,5	14,8
2005	171	4,1	8,8	54	18,5	14,8
2006	172	4,1	8,7	55	18,2	16,4
2007	172	4,1	8,7	55	20,0	20,0
2008	173	4,6	0,0	56	25,0	0,0
2009	174	5,2	0,0	56	28,6	17,9
2010	175	4,6	8,0	57	22,8	22,8
2011	178	5,6	11,2	56	26,8	25,0
2012	170	6,5	0,0	56	26,8	0,0
2013	176	7,4	0,0	56	32,1	0,0
2014	176	9,1	14,2	56	32,1	37,5
2015	177	10,2	15,3	57	35,1	42,1
2016	177	12,4	16,9	57	35,1	50,9
2017	178	13,5	16,9	60	40,0	40,0
2018	178	13,5	16,3	56	42,9	42,9

Для того, чтобы обойти эту трудность, мы применили некоторый опосредованный подход. Он состоит в том, что вместо данных о категориях мы использовали данные о тех журналах, которые соответствуют той или иной категории. Это значит, что вместо некоторой категории *WoS* в качестве ее «полноправного представителя» будет выступать совокупность тех журналов, которые этой категории соответствуют. При таком подходе все вычисления и преобразования, которые необходимо выполнить для определения динамики значений некоторого показателя категории, производятся над данными ее журналов-представителей, тогда как собственно данные указанной категории при опосредованном подходе, ввиду их недостаточности, в этих целях не используются.

ФОРМИРОВАНИЕ СОВОКУПНОСТЕЙ (МНОЖЕСТВ И ПОДМНОЖЕСТВ) ЖУРНАЛОВ

Рассматриваемые множества журналов, представленных в ежегодных выпусках *JCR*, в определенном смысле являются представителями всей мировой науки. С одной стороны, это позволяет утверждать, что те данные о тенденциях значений показателей периода полужизни, которые получатся на основе анализа этих множеств, будут в максимальной степени представительными и надежными. Однако следует отметить и следующее. Эти множества представляют конгломерат журналов. Действительно, каждый ежегодный выпуск *JCR* включает журналы из многих десятков стран (более 80) и соответствует сотням (более 250) тематических категорий *WoS*. Можно предположить, что тенденции, характерные для одного направления исследований (тематической категории *WoS*), наложатся на тенденции, характерные для других направлений исследований и, возможно, взаимно исказят (затемнят) друг друга. То же можно предположить и в отношении региональной составляющей: тенденции значений показателей полупериодов жизни для подмножества журналов, соответствующего некоторой стране, могут отличаться от тенденций для соответствующего подмножества журналов другой страны. Более того, можно ожидать, что такое различие по региональному признаку (стран издания) будет наблюдаться даже внутри одной той же категории *WoS*.

Можно также выделить те подмножества журналов, которые присутствуют в *JCR* на протяжении всего 22-х летнего периода. Тенденции такого «ядра» журналов могут отличаться от тенденций, характерных для всего соответствующего множества журналов. Учитывая все это, необходимо исследовать как тенденции, характеризующие все множество журналов, так и тенденции, характерные для отдельных его подмножеств. Эти подмножества будут сформирова-

ны с использованием следующих классификационных признаков:

a) принадлежность журнала к заданному (i-му) ежегодному выпуску (году опубликования) *JCR*;

b) принадлежность журнала к определенному тематическому выпуску *JCR* (к *JCR-SE* или к *JCR-SSE);*

c) принадлежность журнала к определенной тематической категории WoS и/или к заданному набору этих категорий;

d) принадлежность журнала той или иной стране издания;

e) факт присутствия журнала во всех без исключения ежегодных выпусках *JCR* за заданный период (в нашем случае за 1997–2018 гг.). В этом случае для нас не важно, в каких тематических выпусках (в *JCR-SE* или в *JCR-SSE*) присутствует журнал: важно только, чтобы он был в *JCR* в каждом году. Эти журналы назовем согласно работам [2, 3] *«Абсолютно сохранившимися журналами»* или *«Всегда присутствующими журналами* "*Absolutely preserved journals* or *always present journals (AP Journals)" or* "*Absolutely retentive journals (AR Journals)"*. Иногда, для краткости, мы их будем называть «постоянными». В качестве исходной точки на шкале времени в этом случае примем 1997 г. Можно считать, что такие журналы являются некоторым ядром мировых научных журналов.

Таким образом, сформулированную в начале статьи задачу следует уточнить и расширить, переформулировав ее следующим образом. Необходимо определить динамику каждого из двух показателей (*CdHL* и *CgHL*) путем сопоставления их средневзвешенных значений за определенные годы. Указанное сопоставление должно быть выполнено как для всего рассматриваемого множества журналов, так и для подмножеств журналов, соответствующих наиболее крупным областям знания (естественные, точные, технические, общественные и гуманитарные науки). Кроме того, следует определить динамику показателей для отдельных разделов этих областей (более 250 тематических категорий *Web of Science*), а также для определенных объединений этих категорий: физика, химия, биология, медицина, технические и сельскохозяйственные науки, история, философия и т.п. Важно также установить влияние региональной составляющей (страны) для каждого из рассматриваемых множеств и подмножеств журналов.

Замечание. В дальнейшем, в целях упрощения изложения, в тех случаях, когда это не будет вызывать путаницы, мы будем упоминать только показатель *CdHL*, полагая при этом, что все соображения, определения и вычисления, которые приводятся далее в отношении этого показателя, в полной мере касаются и показателя *CgHL*.

МЕТОДИКА РАСЧЕТА СРЕДНЕВЗВЕШЕННЫХ ЗНАЧЕНИЙ ПОКАЗАТЕЛЕЙ ПОЛУЖИЗНИ

Перейдем к обсуждению конкретных возможностей, предоставляющих имеющиеся данные для решения поставленных задач. Очевидно, что попытки определения динамики средневзвешенных значений рассматриваемых показателей, учитывая только те журналы, для которых значения показателей в *JCR* даются в числовой форме, приведут, в лучшем случае, к существенным искажениям. Ниже предлагаются использованные нами два независимых друг от друга метода расчета соответствующих значений, позволяющие избежать таких искажений, по крайней мере, существенно их ослабить. Один из этих методов (упрощенный, более чем приближенный), который можно назвать методом приписывания численных значений каждому такому журналу, у которого вместо конкретного значения показателя в *JCR* указано «>10», дает возможность действительно лишь очень приближенно оценить средневзвешенные значения заданного показателя. Второй – метод виртуализации распределений – является сложным, однако, он обеспечивает получение значительно более точных значений показателей, и тем самым позволяет получить более достоверную картину динамики исследуемых показателей.

Метод приписывания значений показателя

Этот метод состоит в том, чтобы заменить текстовое выражение «>10» (неполные, неточные данные) на числовое, например, на «10,1»[6]. Это значит, что каждому конкретному (здесь важно подчеркнуть – конкретному) журналу, характеризующемуся такими неполными данными значения показателя, «приписывается» числовое значение. Это позволяет при соответствующих вычислениях учитывать также и те журналы, у которых рассматриваемые показатели имеют значения, превышающие 10 лет. К сожалению, этот метод, хотя и дает возможность оценить, в качестве первого приближения, тенденции изменения рассматриваемых показателей, однако при расчете их средневзвешенных значений приводит к очень большим искажениям. Так, при приписывании минимально возможных численных значений (10,1 года) величина показателя будут существенно занижена. Напротив, при выборе достаточно большого числа, например, 15,0, возникает серьезная опасность необоснованного и очень существенного завышения средневзвешенного значения соответствующего показателя. В настоящем исследовании будет использовано минимальное число 10,1. Такой выбор, несмотря на его произвольность, все же позволяет утверждать, что в этом случае, по крайней мере, не будет нарушено неравенство $10{,}1 <= x$, где x – это одно из возможных значений показателя для какого либо журнала из усеченной части распределения.

Метод виртуализации распределений

Прежде чем перейти к непосредственному изложению этого метода, необходимо ввести ряд определений, а также изложить гипотезу, на которой этот метод основывается.

Распределение журналов по значениям показателя. Назовем распределением журналов по значениям показателя *CdHL/CgHL* таблицу, в первой графе (колонке) которой последовательно в порядке возрастания приводятся значения *CdHL/CgHL*, а во второй

[6] Выбор этого минимального значения может быть оправдан тем, что, по крайней мере не будет необоснованного завышения реальных значений того или иного показателя периода полужизни.

графе против каждого значения *CdHL/CgHL* из первой колонки указано число журналов, каждый из которых имеет именно это значение. В дальнейшем для краткости вместо выражения «распределение журналов по значениям показателя» будем писать «распределение», а вместо «распределения журналов по показателю *CgHL*» – «распределение *CgHL*».

Взаимно однотипные распределения. Под взаимно однотипными распределениями подразумеваем распределения журналов, соответствующие одному и тому же показателю. Так, взаимно однотипными распределениями являются различные распределения журналов по значениям показателя *CdHL*. Взаимно однотипны также все распределения журналов по значениям показателя *CgHL*. Важно помнить, что два распределения, одно из которых соответствует показателю *CdHL*, а другое – *CgHL*, не взаимно однотипны. Сопоставление взаимно разнотипных распределений друг с другом, как правило, не корректно. Оно необходимо только тогда, когда явным образом уточняется, что ставится задача сопоставления именно разнотипных распределений. Следует иметь в виду, что не всегда корректным будет и сопоставление взаимно однотипных распределений. Дело в том, что целесообразно, как правило, сопоставлять друг с другом только те взаимно однотипные распределения, которые принадлежат к одному и тому же классу распределений.

Классы распределений. Класс распределений – множество *однотипных распределений*, полученных на таких подмножествах журналов, каждое из которых сформировано с помощью одного и того же набора классификационных признаков. Так, например, в состав одного и того же класса входят все такие распределения, которые соответствуют показателю *CdHL* и при этом сформированы на основе признака «Страна издания». Еще два класса образуют распределения, причем одни из них соответствуют показателю *CdHL*, а другие – *CgHL* и, при этом полученные на подмножествах журналов, сформированных на основе признака «Категория *WoS*». Еще один пример класса: все распределения, соответствующие показателю *CgHL* и при этом полученные на подмножествах журналов, которые сформированы на основе двух признаков: «Категории *WoS*» **и** «Страна издания» (здесь «**и**» играет роль конъюнкции). Класс может включать набор распределений по *CdHL*, сформированных путем выделения журналов, соответствующих, например, категории «*Psychiatry*» и, в свою очередь, распределённых по странам издания этих журналов. Таким образом, в этот класс входят, в частности, распределения журналов по *CdHL*, которые могут быть описаны следующим образом: «*Psychiatry – Germany*», «*Psychiatry – Russia*», «*Psychiatry – USA*» и т.д.

Группы журналов в распределении по значениям заданного показателя (группы журналов). Для наших целей группа журналов – такая совокупность журналов в некотором распределении, у каждого из которых (журналов) значения заданного показателя численно равны друг другу. Таким образом, в распределении насчитывается столько групп, сколько в этом распределении насчитывается численно различающихся значений этого показателя: от минимального до его максимального значения. Например, группу в некотором распределении *CdHL* составляют журналы, у каждого из которых значение показателя CdHL равно 7,8 года.

Важно учитывать следующее обстоятельство. При равенстве значений двух различных показателей (*CdHL* и *CgHL* соответственно) количество и состав журналов в группе из распределения по показателю *CdHL* не обязательно совпадает (точнее, обычно не совпадает) с количеством и составом журналов в группе из распределения по показателю *CgHL*.

Замечание. Результаты предыдущего этапа нашего исследования изложены в работе [2], в которой использовались понятия «основная часть распределения» и «хвост распределения», причем за «точку раздела» было принято значение показателя, равное 10 годам. При этом за основную часть распределения признавалась та его часть, значения показателя у которой меньше либо равно 10. Соответственно, хвост распределения – та его часть, значения показателя у которой больше 10. Такие определения были «подсказаны» ситуацией, согласно которой почти во всех ежегодных выпусках в качестве максимального числового значения фигурирует именно число 10, а для значений, больших 10, указывается, как мы уже неоднократно отмечали, текстовое выражение вида «>10». К сожалению, предпринятые нами в дальнейшем попытки руководствоваться этими определениями показали их недостаточность. В итоге принято решение отказаться от такого деления распределения. Вместо этого, после серии экспериментов было приняты следующие определения, которыми мы руководствовались в настоящей работе при соответствующих вычислениях.

Части распределения. Рассмотрим распределение по некоторому показателю. Отложим по оси абсцисс значения показателя, т.е. численные обозначения групп журналов. В качестве точки, которая делит распределение на основную часть и хвост, примем значение показателя, соответствующее медиане распределения. Совокупность групп, которая находится слева от медианы, вместе с группой, которая соответствует медиане, будем называть *основной частью распределения*. Совокупность групп, которые расположены справа от указанной точки назовем *хвостом распределения*.

Вычисления, выполненные с целью уточнения алгоритмов и отладки программных средств, показали, что их точность возрастает с делением на более мелкие части. В настоящей работе мы остановились на делении каждой из частей распределения еще на две части. Казалось бы, что в этой ситуации можно отказаться от терминов «основная часть распределения» и «хвост распределения». Действительно, эти термины в нашем случае представляют собой некоторый анахронизм, они иллюстрируют только развитие подходов к решению задач настоящего исследования и связаны с историй этого исследования. Тем не менее, мы не отказываемся от этой терминологии, так как она может оказаться полезной в будущем при детальном анализе исследуемых распределений с точки зрения их типологии и статистических характеристик.

Теоретическое обоснование и описание метода виртуализации

Анализ данных, выполненный ранее в рамках нашего исследования в работе [2], показал, что структуры взаимно-однотипных распределений журналов для всех ежегодных выпусков *JCR* достаточно близки. А именно, доля, которую занимает такая группа в распределении, полученном на массиве журналов некоторого ежегодного выпуска *JCR*, близка к доле соответствующих (по значению заданного показателя) групп остальных ежегодных выпусков. Это оказалось справедливым и для долей более крупных частей распределений. На основании этих и некоторых дополнительных данных в работе [2] была предложена рабочая гипотеза, которую здесь, после некоторого ее уточнения, назовем и сформулируем следующим образом.

Гипотеза подобия распределений, принадлежащих одному и тому же классу. Распределения журналов, принадлежащих одному и тому же классу, по своей структуре подобны друг другу. При этом, чем больше журналов в каждом таком распределении и чем больше доля, которую занимает та часть распределения, в которой находится заданная группа журналов, тем выше вероятность попадания журналов в эту группу, т.е. тем больше журналов окажется в этой группе. Другими словами, число и доля журналов в некоторой группе распределения (положительно) зависят от общего числа журналов в этом распределении, а также находятся в некоторой положительной зависимости от того, какую долю от общего числа журналов занимает та часть распределения, в которой находится эта группа.

На основании этой гипотезы в работе [2] был предложен относительно несложный математический аппарат и, в свою очередь, *функция преобразования неполного (усеченного) распределения в распределение полное*[7]. С помощью этой функции из исходного усеченного распределения создается некоторое виртуальное распределение, которое мы будем рассматривать в качестве полноправного представителя исходного. Такая виртуализация осуществляется с использованием соответствующих данных некоторого полного (реперного) распределения, которое должно принадлежать к тому же классу, к которому принадлежит преобразовываемое распределение. В качестве реперных используются распределения, соответствующие 2018 г., которые, как неоднократно отмечалось выше, в случае журналов всегда являются полными. В качестве исходных данных в функции преобразования используются:

a) данные, которые в самом общем виде описывают преобразуемое распределение: общее число журналов в этом распределении, число журналов в соответствующей части распределения, число журналов, значения показателя у которых больше 10 лет, число журналов с полным отсутствием данных о значениях показателя;

b) данные из реперного распределения, аналогичные тем, которые приведены в (*a*), а также данные о детальной структуре этого (реперного) распределения, а именно: количество (число) журналов, которое соответствует тому или иному значению показателя в реперном распределении.

Помимо этого, в функции преобразования учитывается также число журналов, у которых полностью отсутствуют данные о значениях соответствующего показателя как в преобразовываемом (усеченном) распределении , так и в реперном (полном) распределении, а также вычисляется разность между этими числами, которая затем делится между соответствующими частями преобразуемого распределения. Это деление производится пропорционально доле, которая занимает та или иная часть преобразуемого распределения. Это значит, что такая численная «прибавка» включается в число журналов соответствующей части усеченного распределения.

Для того, чтобы в общем виде представить указанную функцию преобразования, введем следующие обозначения:

G – реперное (полное) распределение, *I* – преобразуемое усеченное распределение, I_{virt} – виртуальное распределение, получаемое с помощью итерационного применения функции преобразования $f(n_G^{d_j} \to n_{I_{virt}}^{d_j})$ и являющееся заместителем (представителем) исходного усеченного распределения *I*;

j – некоторая часть соответствующего распределения;

$n_G^{d_j}$ – число журналов в группе d_j^G, расположенной в части j_G распределения *G*;

$n_{I_{virt}}^{d_j}$ – число журналов в группе $d_j^{I_{virt}}$ расположенной в части $j_{I_{virt}}$ виртуального распределения I_{virt}, которое (число журналов) вычисляется с помощью функции преобразования $f(n_G^{d_j} \to n_{I_{virt}}^{d_j})$ и, требуя при этом, чтобы группа $d_j^{I_{virt}}$ виртуального распределения I_{virt} была численно равна (по значению рассматриваемого показателя) группе d_j^G реперного *G*, т.е. при условии $d_j^{I_{virt}} = d_j^G$;

$\alpha_{I/G}^N$ – отношение [общего числа журналов N_I в усеченном (преобразуемом) распределении *I*] к [общему числу журналов N_G в реперном распределении *G*];

$\mathrm{Re}l_{I/G}^{n_j}$ – отношение [числа журналов n_I^j (с учетом «прибавки», указанной в предыдущем абзаце), находящихся в данной части j_I распределения *I*] к [числу журналов n_G^j в аналогичной части j_G распределения *G*].

В итоге в общем виде функции преобразования может быть записана следующим образом:

$$f(n_G^{d_j} \to n_{I_{virt}}^{d_j}) = \alpha_{I/G}^N * \mathrm{Re}l_{I/G}^{n_j} * n_G^{d_j}$$

[7] Несмотря на то, что математический аппарат, включая функцию преобразования, был разработан, исходя из неуточненной формулировки гипотезы, этот аппарат, как оказалось, полностью соответствует уточнённой формулировке этой гипотезы.

Таблица 4

Динамика средневзвешенных значений *Cited Half-life*, объединенных массивов журналов (*JCR-SE* + *JCR-SSE*)

Год выпуска *JCR*	Для всех журналов			Для абсолютно сохранившихся журналов		
	Способ полной виртуализации	Способ частичной виртуализации	Способ приписывания значений	Способ полной виртуализации	Способ частичной виртуализации	Способ приписывания значений
	Период полужизни, годы (тысячные доли)					
1	**2**	**3**	**4**	**5**	**6**	**7**
1997	7,999	7,695	6,569	10,422	8,874	6,526
1998	8,056	7,830	6,640	10,450	8,955	6,667
1999	8,025	7,769	6,652	10,457	8,952	6,750
2000	8,031	7,787	6,668	10,423	8,963	6,839
2001	8,033	7,781	6,689	10,391	8,994	6,956
2002	8,057	7,790	6,700	10,331	9,023	7,055
2003	8,025	7,737	6,697	10,325	9,034	7,124
2004	8,000	7,734	6,749	10,293	9,099	7,231
2005	8,003	7,718	6,732	10,267	9,120	7,298
2006	8,047	7,739	6,750	10,258	9,161	7,402
2007	8,031	7,762	6,758	10,235	9,207	7,501
2008	7,992	7,726	6,782	10,197	9,274	7,616
2009	7,958	7,754	6,814	10,225	9,406	7,800
2010	8,010	7,707	6,680	10,241	9,424	7,828
2011	8,002	7,698	6,677	10,265	9,538	7,933
2012	8,000	7,754	6,732	10,353	9,676	8,057
2013	7,963	7,724	6,801	10,397	9,831	8,187
2014	7,955	7,759	6,884	10,505	9,994	8,313
2015	7,903	7,776	6,975	10,589	10,160	8,439
2016	7,820	7,776	7,124	10,676	10,367	8,608
2017	8,541	8,480	7,388	10,757	10,757	8,805
2018	8,541	8,541	7,415	10,926	10,926	8,845

Уточним, что такие вычисления должны быть последовательно выполнены для каждой группы журналов, образующих реперное распределение. Если реперное распределение G состоит, например, из S групп, то для получения соответствующего виртуального распределения следует осуществить S итераций таких вычислений. В итоге вместо исходного неполного (усеченного) распределения I мы получим некоторое виртуальное, но теперь уже полное распределение I_{virt}, включающее все те группы (значения показателя), которые содержит реперное распределение G, причем для каждой из этих групп виртуального распределения будет указано соответствующее ей вычисленное число журналов. Полученное таким образом виртуальное распределение I_{virt} в дальнейших расчетах рассматривается в качестве полноправного представителя (заместителя) исходного усеченного распределения I. Средневзвешенное значение показателя для данного исходного распределения I вычисляется уже не непосредственно на основе данных этого распределения, а опосредовано, т.е. на основе данных того виртуального распределения I_{virt}, которое построено как представитель исходного.

Следует отметить, что метод виртуализации и разработанные для его реализации программные средства позволяют для одного и того же распределения осуществлять вычисления средневзвешенного значения соответствующего показателя двумя различными, но частично совпадающими способами.

А. *Способ полной виртуализации*. На основании гипотезы подобия и с использованием функции преобразования для усеченного распределения предварительно строится соответствующее ему виртуальное распределение. Это значит, что из текущего (усеченного) распределения используются только следующие данные: общее число журналов, число (доля) журналов в соответствующих частях распределения, а также число журналов, у которых полностью отсутствуют данные о значениях показателя. Важно понимать, что при этом игнорируются все значения показателя у всех журналов, которые формируют исходное распределение, даже в тех случаях, когда реальное значение показателя <=10 годам.

Таким образом, виртуальное распределение, которое теперь является «заместителем» исходного усеченного, полностью строится только с помощью перерасчета соответствующих значений реперного (полного, не усеченного) распределения, конечно, с использованием тех данных из исходного распределения, которые перечислены в предыдущем абзаце. После построения виртуального распределения вычисляется сумма произведений каждого значения показателя этого распределения на соответствующее ему число журналов, а затем полученная сумма делится на общее число журналов в распределении. Результат этого деления представляет средневзвешенное значение соответствующего показателя и рассматривается нами в качестве представителя той совокупности журналов, которые образовали исходное распределение. Например, если исходное распределение образовано из журналов выпуска *JCR* 2010 г., то полученное таким способом значение *CdHL*, которое равно 8,011 (графа 2 табл. 4) характеризует эту совокупность журналов по состоянию именно в 2010 г.

В. С*пособ частичной виртуализации*. В этом случае преобразование исходного (усеченного) распределения осуществляется только для его усеченной части, т.е. той части распределения, которой соответствуют текстовые выражения вида «>10». Что касается части, у которой показатель принимает численные значения (т.е. не превышающие 10 лет), то она преобразованию не подвергается. Из этих двух частей составляется результирующее распределение, являющееся некоторым «склеенным» (гибридным) распределением, состоящим из двух частей: одна часть – это неизменённый фрагмент исходного распределения со значениями показателя <=10, другая – результат преобразования того фрагмента этого распределения, значения показателя которого превышает 10 лет. Средневзвешенное значение соответствующего показателя в этом случае вычисляется на основе данных такого склеенного (гибридного) распределения.

ДОСТОИНСТВА И НЕДОСТАТКИ ПРЕДЛОЖЕННЫХ СПОСОБОВ ВЫЧИСЛЕНИЯ РАСПРЕДЕЛЕНИЙ

В результате применения двух предложенных методов для одного и того же распределения по некоторому заданному показателю можно получить три различных значения средневзвешенного показателя:

1) средневзвешенное значение, полученное с помощью метода приписывания значений;

2) средневзвешенное значение, полученное с помощью метода полной виртуализации;

3) средневзвешенное значение, полученное с помощью метода частичной виртуализации.

Серьёзным недостатком метода приписывания значений является то, что приписываемое значение принимается произвольно, только исходя из того, чтобы оно было больше 10 лет, т.е. как этого требует заменяемое выражение («>10»). При этом принятое в настоящей работе минимально возможное значение (10,1 года) может заметно занижать вычисляемые таким образом значения. Достоинство этого метода – относительно несложные операции вычисления, хотя в случае использования компьютеров это оказывается не столь существенным. В любом случае, при помощи метода приписывания значений можно определить тенденцию в изменениях этих значений, если таковая имеет место, либо, напротив, убедиться в отсутствии каких-либо изменений.

Достоинство метода полной виртуализации состоит, прежде всего, в том, что он позволяет сопоставить исходное и виртуальное распределение, оценить степень их близости и, тем самым, дать возможность оценить насколько справедлива гипотеза подобия. Что касается недостатков, то здесь следует отметить, что вычисление средневзвешенного значения на основании данных только виртуального распределения полностью игнорирует реальные данные исходного и, тем самым, скорее всего, может существенно искажать результат. Очевидно, что наиболее точными окажутся результаты, полученные с помощью метода частичной виртуализации, так как в этом случае, с одной стороны, – используются все реальные данные исходного распределения, а, с другой, – с помощью виртуализации восполняются недостающие данные.

РЕЗУЛЬТАТЫ ИССЛЕДОВАНИЯ

Сформируем массивы журналов по следующей схеме:

a) из тематического выпуска *JCR-SE* и тематического выпуска *JCR-SSE* для каждого ежегодного выпуска *JCR* сформируем объединенный массив уникальных журналов, т.е. журналов, неповторяющихся в рассматриваемом году. Это значит, что в такой массив попадут все журналы, которые в указанном году присутствовали хотя бы в одном из двух тематических выпусков *JCR*;

b) потребуем, чтобы журнал в обязательном порядке присутствовал в каждом ежегодном массиве, сформированном на предыдущем этапе. В итоге получим ежегодные массивы журналов, которые обычно мы обозначаем как абсолютно сохранившиеся журналы.

График средневзвешенных значений показателя *CdHL* именно для сформированных таким образом массивов абсолютно сохранившихся журналов приведен на рис.1. Эти значения получены тремя различными способами. Первый (способ полной виртуализации) полностью базируется на гипотезе подобия. Второй (способ частичной виртуализации) также использует гипотезу подобия, однако только для восполнения недостающих данных. И наконец, третий способ (способ приписывания значений) никак не обращается к гипотезе подобия. Следовательно, для ответа на вопрос, справедлива ли гипотеза подобия и если справедлива, то в какой степени и в каких пределах, нам нужно рассмотреть результаты, полученные с помощью двух первых способов.

Из графика на рис. 1 и соответствующих ему граф 5–7 табл. 4 следует, что результаты, полученные с помощью метода, который полностью опирается на гипотезу подобия (способ полной виртуализации), достаточно близки к тем результатам, которые получены исходя из указанной гипотезы подобия, но при обязательном использовании реальных (неусеченных) данных. Здесь важно подчеркнуть, что доля

этих данных в усеченных распределениях очень значительна (см. графу 6 табл. 1). Следовательно, можно утверждать, что полученные результаты не противоречат гипотезе подобия, а на отрезках времени в 8 – 10 лет средневзвешенные значения показателя *CdHL*, полученные с помощью метода, полностью базирующегося на гипотезе подобия (метод полной виртуализации), достаточно хорошо описывают динамику этого показателя. Таким образом, мы можем не только констатировать, что результаты не противоречат гипотезе подобия, но и указать те временные рамки, в которых справедливость этой гипотезы не вызывает сомнений.

В качестве более точной оценки справедливости гипотезы подобия можно было бы использовать разность между средневзвешенным значением заданного показателя (*CdHL* или *CgHL*), вычисленным для данного полного распределения (например, некоторого годового выпуска *JCR*) – с одной стороны, и значением аналогичного показателя, рассчитанного для этого же распределения, но на основе гипотезы подобия – с другой. К сожалению, возможности такой оценки в настоящее время ограничены. Действительно, единственный год выпуска *JCR*, когда можно выполнить такие сопоставительные вычисления – это 2017 г.: только в выпуске 2017 г. (как и в реперном выпуске 2018 г.) отсутствует ситуация, когда вместо конкретного значения показателя указано «>10». Выполненные нами вычисления для указанного годового выпуска *JCR* дали следующие значения: 10,725 (по реальным значениям) и 10,757 (способ полной виртуализации). Разность между этими вычислениями составила – 0,032 года (0,29%). Такая разность соответствует временному расстоянию между реперным (2018) и заданным (2017) годом, которое равно одному году. Если предположить, что указанная разность линейно зависит от величины временного расстояния, то это отклонение, например, для 2007 г. составит – 0,32 года (–0,032*10= – 0,32), что примерно соответствует данным на графике (рис.1). Этот график позволяет также оценить точность и достоверность результатов, полученных с помощью способа приписывания значений. Как и следовало ожидать, и как следует из графика, способ приписывания значений существенно занижает средневзвешенные значения показателя *Cited Half-life*. Тем не менее, полученные с помощью этого метода результаты достаточно хорошо отражают тенденцию изменения во времени этого показателя: с определенной степенью огрубления можно утверждать, что кривая, соответствующая этому способу на графике, эквидистантна кривой, соответствующей способу частичной виртуализации. Анализ графика, продемонстрированного на рис. 1, позволяет заключить, что наиболее точную картину о динамике показателя можно получить с помощью способа частичной виртуализации. Именно этим методом мы будем пользоваться в дальнейших исследованиях.

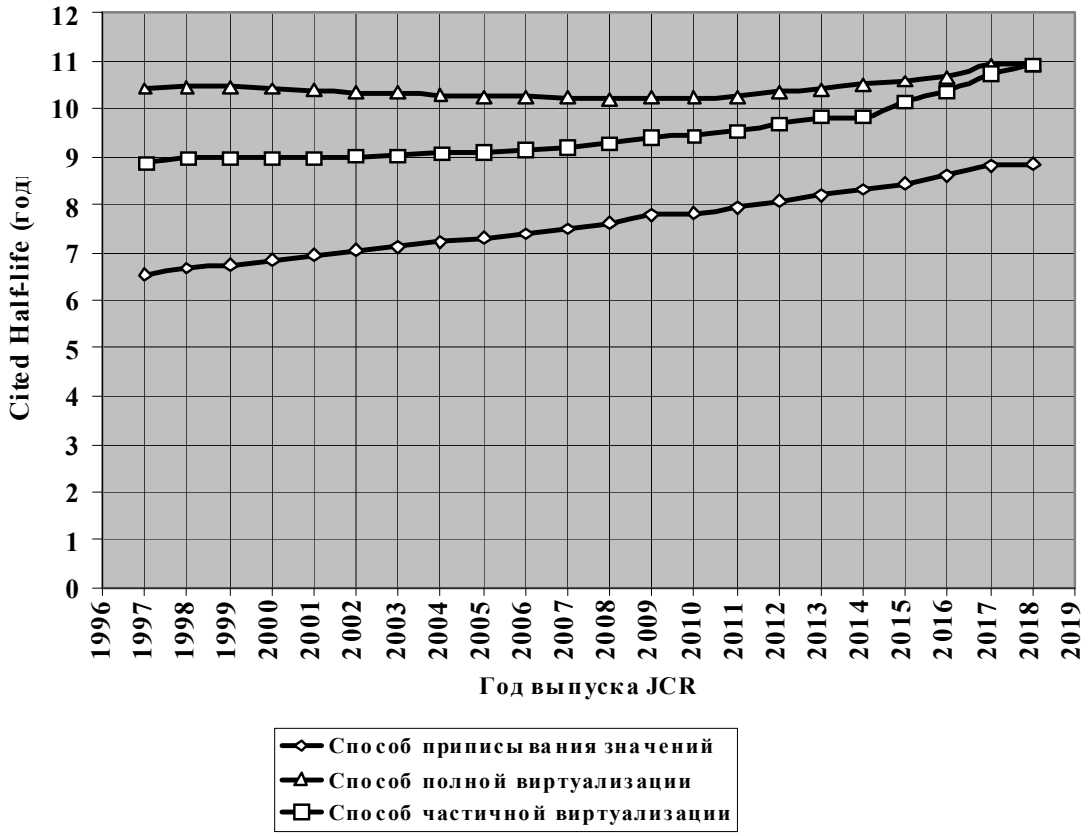

Рис. 1. Сопоставление значений показателя *Cited Half-Life*, полученных с помощью различных способов расчета (абсолютно сохраняющиеся журналы *JCR*)

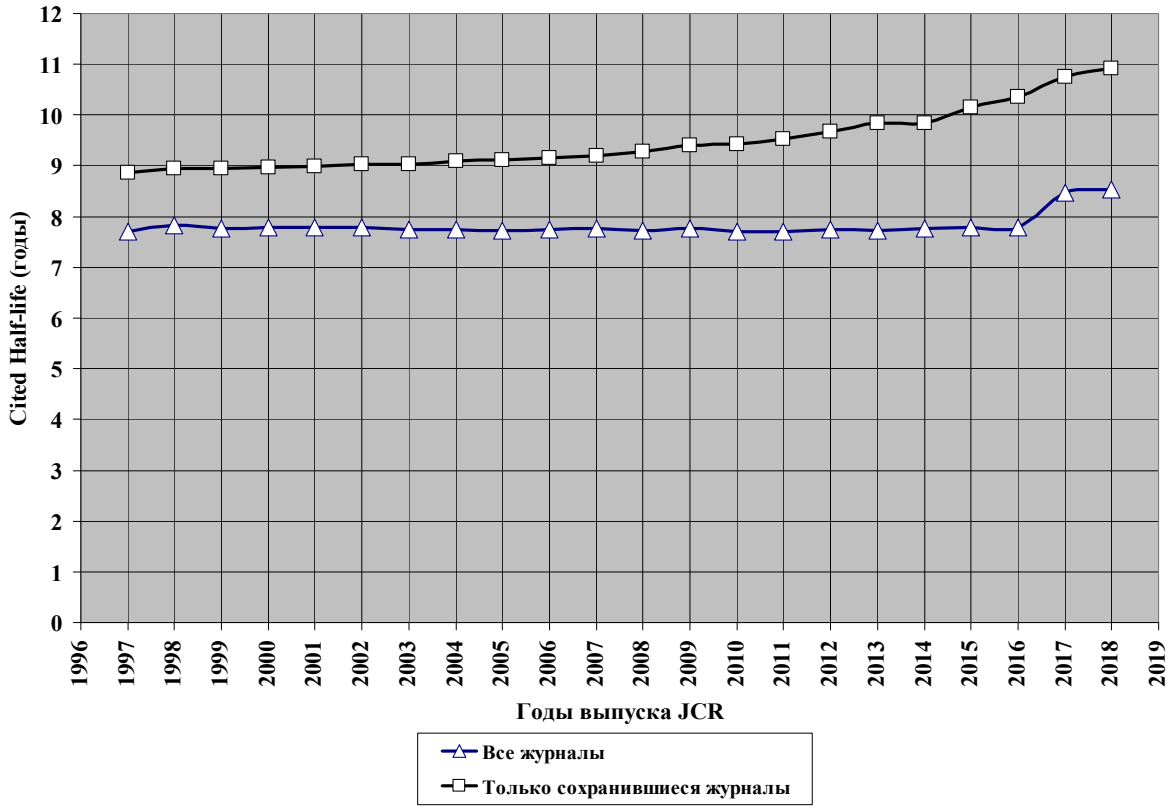

Рис. 2. Сравнение значений показателя *Cited Half-life* для массивов журналов *JCR* и их подмассивов, содержащих только абсолютно сохранившиеся журналы (результаты, получены с помощью способа частичной виртуализации)

Несколько важных выводов позволяют также сделать табл. 4 и соответствующий ей график, показанный на рис. 2. Средневзвешенные значения показателя *CdHL* для массивов, состоящих из «абсолютно сохраняющихся журналов», т.е. массивов, сформированных с учетом требований (*a*) и (*b*) всегда больше, чем соответствующие значения для массивов, сформированных только с учетом требования (*a*) – «все журналы». Кроме того, следует отметить, что изменения значений этого показателя в случае «абсолютно сохраняющихся журналов» происходят более динамично, чем соответствующие изменения в случае «все журналы». Более того, начиная с 1999 г. и до 2011 г. в случае «все журналы» динамика имеет очень слабую (едва заметную) отрицательную тенденцию. Затем тенденция медленно меняется на противоположную, и только в последние два года значение показателя начинает быстро расти.

ЗАКЛЮЧЕНИЕ

Уточнена ранее сформулированная нами гипотеза подобия распределений журналов по значениям показателей Cited Half-life Citing Half-life. На большом эмпирическом материале осуществлена проверка этой гипотезы, которая подтвердила ее справедливость. Установлено, что гипотеза выполняется для тех пар взаимно-однотипных распределений, которые принадлежат одному и тому же классу (подклассу) распределений. Установлено, что в пределах 8-10 лет гипотеза выполняется с большой точностью, однако и в интервале 15-20 лет ее использование для оценки динамики показателей *CdHL* и *CgHL* оказывается достаточно эффективным.

Скорректированы понятия основной части и хвоста распределения. Предложено и применено дополнительное деление этих частей распределения, что позволяет вычислять средневзвешенные значения показателей *CdHL и CgHL* с существенно большей точностью. Детально проработаны методы определения средневзвешенных значений соответствующих показателей. Разработаны средства, с помощью которых сформированы многие десятки классов взаимно-однотипных распределений, суммарно включающих тысячи распределений. Эти средства также позволяют реализовать предложенные методы на указанных распределениях.

Установлено, что для случаев «абсолютно сохранившихся журналов» значения показателя *CdHL* характеризуются достаточно ярко выраженной положительной динамикой, тогда как для случаев «все журналы» о положительной динамике можно говорить только в пределах временного отрезка, соответствующего последним трем годам.

Таким образом, созданы алгоритмы и методы вычисления, на основе которых работают программы, позволяющие наиболее точно определять период по-

лужизни научных журналов и их совокупностей, объединенных тематическими категориями *JCR*. В конечном счете это дает возможность сравнивать динамику развития областей науки, отраслей знания и отдельных научных дисциплин. Поскольку среди способов планирования и прогнозирования, применяемых в самых различных областях науки, самым традиционным и проверенным является экстраполяция тенденций развития, период полужизни научной литературы будет служить для этой цели важным подспорьем. Он также может применяться и при отслеживании долговечности идей выдающихся ученых [4]. Особенно важно, что предложенные нами способы впервые позволяют проследить эти тенденции по периоду полужизни не только цитируемых, но и цитирующих статей. Между этими показателями существует определенная, но пока еще не изученная корреляция, которая заслуживает отдельного анализа.

СПИСОК ЛИТЕРАТУРЫ

1. Burton R.E., Kebler R.W. The “half-life” of some scientific and technical literature // American Documentation. – 1960. – Vol. 11, № 1. – P. 18-22.
2. Либкинд А.Н., Маркусова В.А., Либкинд И.А. К вопросу определения динамики показателей периода полужизни журналов по *Journal Citation Reports* // Научно-техническая информация. Сер. 2. – 2020. – № 5. – С. 29-38; Libkind A.N., Markusova V.A., Libkind I.A. Approach for Using Journal Citation Reports in Determining the Dynamics of Half-Life Indicators of Journals //Automatic Documentation and Mathematical Linguistics. – 2020. – Vol. 54, № 3. – P. 174-184.
3. Либкинд А.Н, Маркусова В.А., Либкинд И.А., Янц М., Иванов К.Н. Моделирование динамики процесса сохранения журналов в качестве наиболее авторитетных научных изданий // Научно-техническая информация. Сер. 2. – 2013. – № 3. – С. 9-34; Libkind A.N., Markusova V.A., Libkind, I.A. Jansz M., Ivanov K.N. Modeling the dynamics of the retentivity process of journals among the most authoritative scientific serials // Automatic Documentation and Mathematical Linguistics. – 2013. – Vol. 47, № 2. – P. 69-92.
4. Розенберг Г.С. «Хиршивость» науки и период полураспада цитируемости научных идей // Биосфера. – 2018. – Т. 10, № 1. – С. 52-64.
5. Москалева О.В. Использование наукометрических показателей для оценки научной деятельности // Науковедческие исследования. – 2013. – С. 85-109.
6. Liang G., Hou H., Chen Q. et al. Diffusion and adoption: an explanatory model of “question mark” and “rising star” articles // Scientometrics. – 2020. – Vol. 124. – P. 219-232.

Материал поступил в редакцию 05.07.2020

Сведения об авторах

ГИЛЯРЕВСКИЙ Руджеро Сергеевич – доктор филологических наук, профессор, главный научный сотрудник, заведующий отделением ВИНИТИ РАН; профессор факультета журналистики Московского государственного университета им. М.В. Ломоносова
e-mail: ruggero29@gmail.com

ЛИБКИНД Александр Наумович – кандидат технических наук, ведущий научный сотрудник ВИНИТИ РАН, Москва
e-mail: anliberty@mail.ru

БОГОРОВ Валентин Григорьевич – руководитель отдела образовательных программ *Clarivate Analytic*s, Москва
e-mail: valentin.bogorov@clarivate.com

ЛИБКИНД Илья Александрович – ведущий аналитик, ООО Сервисное бюро ВИП, Москва
e-mail: Libkind_Ilya@hotmail.com

УДК [004.75.056:004.41]:005.92

А.Ю. Щербаков

Комплексный подход к созданию платформы доверенного документооборота с электронной подписью

Рассматриваются возможности создания платформы доверенного документооборота с электронной подписью, предназначенной для взаимной проверки и признания документов, сформированных в различных юрисдикциях и снабженных электронной подписью с использованием различных алгоритмов и ключей. Представлен способ надежного доверенного хранения, обмена и проверки документов на основе доверенного распределенного реестра с использованием смарт-контрактов и симметричных криптографических алгоритмов.

Ключевые слова: *код аутентификации, распределенный реестр, блокчейн, протокол, ключи, электронная подпись, средства криптографической защиты информации, безопасность, датчик случайных чисел, удостоверяющий центр, смарт-контракт*

DOI: 10.36535/0548-0027-2020-11-3

ВВЕДЕНИЕ

Задача взаимной проверки электронных подписей (ЭП)[1], сформированных по различным алгоритмам в различных юрисдикциях, является сегодня весьма актуальной и, к сожалению, пока технически нерешенной.

Предположим, что имеется несколько участников системы, находящихся в различных государствах (юрисдикциях) и имеющих различные алгоритмы и ключи электронной подписи, для которых необходима общая возможность подтверждения документов, подписанных этими подписями [1]. Классическое решение этого вопроса, связанное с созданием единого доверенного удостоверяющего центра, практически невозможно, поскольку затруднительно выработать общие технические и юридические регламенты работы центра. Кроме того, в силу закрытости процедур оценки качества криптографических механизмов, участники системы находятся в весьма затруднительном положении при опубликовании результатов криптографических исследований качества электронных подписей.

При этом мы понимаем, что оценке надежности подвергаются как совокупность алгоритмов ЭП, так и свойства удостоверяющего центра, который формирует сертификаты – подписывает своей электронной подписью открытые ключи пользователей (участников системы). Кроме того, сами процедуры по проверке ЭП в системе часто проводятся автоматизировано – с участием смарт-контрактов.

Известно решение, называемое «служба доверенной третьей стороны» (служба ДТС – *Litoria DVCS*) [2], которое, по мнению его авторов, обеспечивает юридическую значимость квалифицированной электронной подписи на иностранном алгоритме в юрисдикции Российской Федерации, проверяя эту иностранную подпись и выдавая юридически значимую квитанцию с результатами проверки. Юридическая значимость квитанции обеспечивается электронной подписью, выполненной с помощью отечественного криптографического алгоритма.

1. Иностранная электронная подпись передается на проверку в службу доверенной третьей стороны.

2. ДТС определяет криптографические алгоритмы, на которых выполнена ЭП и передает эту электронную подпись в ДТС национального сегмента, который может легитимно работать с данными криптографическими алгоритмами.

3. ДТС национального сегмента проверяет электронную подпись, **создает квитанцию с результатами проверки** и передает квитанцию в ДТС сегмента Российской Федерации (ДТС РФ).

4. ДТС РФ подписывает квитанцию отечественной электронной подписью, тем самым придавая юридическую значимость иностранной ЭП, результаты проверки которой зафиксированы в квитанции.

5. ДТС РФ возвращает подписанную квитанцию пользователю.

Очевидно, что предлагаемая схема имеет слабое место в п.3, когда квитанция передается в ДТС РФ,

[1] Федеральный закон "Об электронной подписи" от 06.04.2011 N 63-ФЗ. – URL: http://www.consultant.ru/document/ cons_doc_LAW_112701/

а проверить электронную подпись под ней не представляется возможным.

Далее мы покажем, что применение распределенного реестра в совокупности с доверенным смарт-контрактом позволит решить эту проблему.

Электронная подпись включает три криптографических алгоритма: 1) **алгоритм хеширования**, который преобразует подписываемый текст (документ) в вектор фиксированной длины (результат хеширования); 2) **алгоритм подписания (простановки ЭП)**, оперирующий с результатом хеширования, секретным (приватным) ключом пользователя и некоторым случайным числом; 3) **алгоритм проверки ЭП**, оперирующий с результатом хеширования и открытым ключом (сертификатом) пользователя.

РЕШЕНИЕ ЗАДАЧИ ВЗАИМНОЙ ПРОВЕРКИ ЭЛЕКТРОННЫХ ПОДПИСЕЙ

Предлагается следующее решение задачи построения платформы документооборота с возможностью проверки различных электронных подписей на базе доверенного распределенного реестра (ДРР), который реализует платформу доверенного документооборота. Доверенность понимается в смысле сохранения свойств документов в системе на протяжении их жизненного цикла. При этом платформа имеет и архивную функцию, поскольку удаление звеньев распределенного реестр (РР) невозможно.

Каждый участник документооборота регистрируется в качестве участника в системе РР и получает возможность размещать и получать документы в этом реестре. При этом целостность и авторство документов проверяется оператором РР при помещении документов в реестр. В распределенном реестре документы помещаются с локальной электронной подписью своего владельца.

Таким образом, в распределенном реестре каждый документ снабжен как подписью владельца, так и кодом аутентификации (КА) РР, который позволяет убедиться в целостности информации и установить ее принадлежность. В качестве кода аутентификации предлагается функция $I=Im(x, k)$ – функция вычисления имитовставки от информации x на ключе k. Эта функция обладает возможностью как авторизации пользователя, так и контроля целостности передаваемой и хранимой информации. Она построена на симметричном криптографическом алгоритме и при этом не является средством криптографической защиты информации, на которое распространяются экспортные и импортные ограничения.

Кроме того, пользователи могут открывать доступ к своим документам другим участникам, а оператор РР может направить документ для проверки электронной подписи тому, кто может выполнить эту операцию, исходя из наличия ключей проверки (сертификатов) ЭП и соответствующих криптографических алгоритмов, и получить квитанцию о верности или неверности ЭП, которая также защищена кодом аутентификации и помещена в распределенный реестр.

Таким образом формируется общая инфраструктура доверия, в которой каждый документ подписан как минимум одной ЭП и кодом аутентификации участника, что позволяет убедиться в подлинности информации и анализировать квитанции о верности ЭП, также защищенные КА. Дополнительно все документы образуют единую цепочку с невозможностью исключения документа из нее, имеют метки времени и сквозную нумерацию. Вся система образует доверенную СУБД с возможностью разветвленного поиска и предоставления различных выборок и статистик в соответствии с правами и ролями участников.

Регламент работы системы, включающий помещение документов в распределенный реестр и запросы на проверку электронной подписи, может быть описан достаточно простым гражданско-правовым договором, который в идентичной форме заключается в каждой юрисдикции и описывает не использование ЭП, а движение документов в системе распределенного реестра.

Введем дополнительные термины для описания дальнейших способов решения задачи создания платформы.

Смарт-контракт (СК) – последовательность команд и вызовов исполняемых модулей из хранилища приложений.

Компилятор смарт-контракта – программный модуль, который переводит команды и имена модулей в сжатую кодированную форму (байт-код).

Хранилище приложений – область данных сервера СК, в которой находятся приложения для выполнения с предварительной проверкой их имитовставки (кода аутентификации) на ключе вызывающего пользователя.

Предположим, что смарт-контракт представляет собой последовательность команд и вызовов исполняемых модулей из хранилища приложений, выполняется на отдельном сервере (сервер смарт-контрактов) и связан по данным с сервером оператора распределенного реестра (сервер оператора РР).

Смарт-контракт формируется пользователем и направляется в систему как стандартный элемент распределенного реестра, описанный выше обычным порядком.

В системе вводятся три вида смарт-контрактов: текстовый, байт-код и анонимизированный байт-код, в последнем вызовы приложений анонимизированы и не видны другим пользователям.

Сервер оператора выделяет смарт-контракт из входного потока данных и направляет его на сервер СК, где он верифицируется и, при положительном результате верификации, исполняется.

Верификация заключается в проверке наличия вызываемых приложений в хранилище приложений и их доступности для данного пользователя, а также в правильном синтаксисе написании команд.

При возникновении ошибки в некоторой строке смарт-контракта сервер СК формирует квитанцию, где указывает код ошибки и строку, в которой она произошла, и направляет эту квитанцию на сервер оператора РР.

После успешного выполнения смарт-контракта также формируется квитанция для пользователя, направившего смарт-контракт в систему.

Приложения, доступные для пользователя, в обязательном порядке подписаны имитовставкой (кодом аутентификации) на его ключе.

КРАТКОЕ ОПИСАНИЕ ТЕХНОЛОГИИ ПЛАТФОРМЫ

Пользователь является абонентом системы РР, оператор которого располагает **модулем доверенного хранения пользовательских ключей** (МДХПК), а подписанные электронной подписью и кодом аутентификации документы хранятся в распределенном реестре [3].

МДХПК при помощи качественного датчика случайных чисел формирует ключи КА, хранит их без выдачи во внешнюю среду (неизвлекаемо) и выполняет на них функции проверки кода аутентификации.

Аутентификация пользователя производится с применением присылаемых на мобильное устройство одноразовых паролей, которые вводятся в программы связи с сервером и используются для реализации протоколов аутентификации. При необходимости на мобильное устройство может быть передан ключ для создания защищенного сеанса с сервером, закрытый на одноразовом пароле.

Понятие «неизвлекаемость» на современном техническом уровне можно трактовать так: в хранилище, понимаемом как изолированное техническое устройство, нет возможности прочитать загруженный или сформированный ключ – по причине отсутствия программных и технических интерфейсов извлечения ключа во «внешний мир». Также невозможно извлечение любой информации о ключе, существенно раскрывающей его содержание, и отсутствует (в техническом плане) информация о ключе в технических каналах наблюдения (электромагнитном, акустическом, визуальном).

Кроме того, модуль доверенного хранения пользовательских ключ должен обеспечивать выполнение криптографических операций (преобразований) на загруженном в него ключе, а также выработку собственных ключей без их извлечения и выдачу вовне только результатов криптографических операций.

При соблюдении свойств «необратимой или сингулярной» загрузки ключей (по аналогии с физической «черной дырой», куда информация и материя попадают безвозвратно) процессы управления ключами осуществляются по их номеру или идентификатору. При этом для хранения идентификаторов возможно применение распределенного реестра, а для верификации данных участников системы, их аутентификации и разрешения споров – использование ключей.

Важнейшим моментом этой технологии являются механизмы качественной генерации ключей, т. е. корректной работы датчиков случайных чисел и проверки качества случайных последовательностей.

Термины и обозначения для формального описания алгоритма платформы

Xi – пользователь корпоративной системы.

NMOi – номер мобильного устройства пользователя.

Ai –информация, описывающая пользователя Xi.

Kxi – персональный ключ пользователя, неизвестный самому пользователю и хранимый в МДХПК.

Ksi – сетевой ключ пользователя (также являющийся частью персональной информации пользователя), предназначенный для связи с оператором РР.

Ci – ключевой контейнер пользователя, представляющий собой персональную информацию пользователя (сетевой ключ), закрытый на пароле пользователя при помощи обратимой криптографической процедуры.

Si – сетевое имя пользователя, однозначно связанное с Ai.

INFOij – информация *i*-го пользователя, сформированная на его рабочем месте и направляемая для хранения и обработки, имеющая условный номер *j* и содержащая ЭП, зависящую от этой информации.

Kvij – квитанция, сообщающая о результате обработки *j*-го информационного блока для *i*-го пользователя.

$I=Im(x, k)$ – функция вычисления имитовставки от информации x на ключе k.

Напомним, что функция вычисления имитовставки обладает возможностью как авторизации пользователя, так и контроля целостности передаваемой и хранимой информации.

Краткое описание команд смарт-контракта

Команда загрузки данных из РР

rload номер_реестра номер_записи файл

где:

номер_реестра – десятичный номер реестра, из которого извлекаются данные,

номер_записи – десятичный номер извлекаемой записи,

файл – имя файла, куда извлекается запись.

Команда возвращает код ошибки или 0 при успешном выполнении.

Команда выгрузки данных в РР

upload файл номер_реестра имя_пользователя номер

где:

файл – имя файла, загружаемого в реестр,

номер_реестра – десятичный номер РР,

имя_пользователя – имя пользователя, которому предназначена формируемая запись,

номер – имя файла, в который при успешном завершении помещается номер сформированной записи (звена РР).

Команда возвращает код ошибки или 0 при успешном выполнении.

Общие команды пакетных файлов

Copy, del, md, cd, ren, dir также поддерживаются в смарт-контракте.

Команда вызова приложений

имя_приложения аргументы,

где:

имя_приложения – имя вызываемого приложения из хранилища приложений,

аргументы – названия аргументов или файлов, с которыми работает приложение.

Команда создания квитанции

mkk файл имя_пользователя,

где:

файл – файл содержащий тело квитанции,

имя_пользователя – имя пользователя, которому предназначена квитанция.

Например,

```
rem Перевод с кошелька k1 пользователя
7d5402af на кошелек пользователя 20ef84c5
rload 1 8146 k1
rload 1 9378 k2
perevod 50 k1 k2
upload k1 1 7d5402af num1
upload k2 1 20ef84c5 num2
```

В приведенном примере смарт-контракта использовано приложение-перевод (perevod), которое переводит 50 условных единиц с кошелька k1, загруженного из записи 8146 реестра 1, на кошелек k2, загруженного из записи 9378 также реестра 1, т. е. k2 = k2 + 50, k1 = k1 - 50, после чего записи в реестре обновляются, номера новых записей помещаются в файлы num1 и num2, соответственно. Квитанции в данном случае могут быть сформированы автоматически оператором РР, либо сервером СК.

Полагаем, что пользователь системы имеет персональный вычислитель (ноутбук, смартфон или выделенный криптокомпьютер), подключенный при помощи каналов связи (телекоммуникационной среды) к оператору РР.

Введем следующие обозначения:

y=E(x, k) – алгоритм зашифрования данных x на ключе k,

x=D(y,k) – алгоритм расшифрования данных x на ключе k,

Pi – пароль пользователей для защиты соответствующих контейнеров.

Пользователь формирует запрос Zij = Im([Si,DATAj], Ksi) и направляет его в распределенный реестр. Сервер проверяет имитовставку пользователя по запросу, тем самым проводя как аутентификацию отправителя, так и проверку целостности данных DATAj, в состав которых могут входить и подписанные электронной подписью данные, описанные нами далее.

При положительном результате проверки запроса информация передается в распределенный реестр и модуль доверенного хранения пользовательских ключей. При положительном результате записи обработанных данных в средствах хранения данных для пользователя формируется квитанция Kvij, содержащая номер блока, куда помещена информация пользователя, номер транзакции, время помещения в распределенный реест и имитовставка на Ksi под указанными данными квитанции пользователя.

Описание протокола

Пользователь владеет мобильным устройством, имеющим связь с оператором РР, к которому подключен модуль доверенного хранения пользовательских ключей, формирующий при помощи качественного датчика случайных чисел ключи, хранящий их без выдачи во внешнюю среду и выполняющий на них функции вычисления и проверки кода аутентификации.

Регистрация пользователя

Пользователь при помощи веб-интерфейса – приложения, установленного на его мобильном устройстве, либо посылкой СМС направляет запрос на регистрацию в системе. Для этого он указывает (для СМС эта информация определяется автоматически) номер своего мобильного устройства NMOi. Оператор РР помещает номер пользователя в базу данных, дает команду МДХПК на выработку случайного ключа Ksi для связи пользователя Xi с оператором РР.

Модуль доверенного хранения пользовательских ключей вырабатывает Ksi, проверяет его статистические свойства и при положительном результате проверки запоминает Ksi, экспортирует его в виде контейнера Csi, закрытого на пароле Pi. Затем МДХПК вырабатывает Kxi, проверяет его статистические свойства и при положительном результате проверки запоминает Kxi, а также вырабатывает сетевое имя Si, определяющее идентификаторы ключей Ksi и Kxi, используемое для операций с пользовательскими данными пользователя Xi с именем Si.

Пароль Pi передается пользователю на материальном носителе (карта памяти, помещаемая в мобильное устройство) или присылается по СМС. Ci также может быть записан на карту памяти, либо выслан по открытым каналам, либо помещен во внешнюю (облачную) систему хранения данных.

Пользователь имеет возможность изменить Pi (с изменением контейнера Ci).

Обмен информацией

Пользователь Xi передает информацию INFOk пользователю Xj.

Для этого он зашифровывает INFOik на ключе Ksi (извлеченном из Ci) и передает E(Ksi, INFOik) в РР, помещающий принятую информацию в МДХКП, который, в свою очередь, расшифровывает информацию внутри себя, зашифровывает на Kxi и передает E(Kxi, INFOik) в корпоративную или внешнюю систему хранения (возможно, в корпоративный распределенный реестр или в облако). При помещении E(Kxi, INFOik) в корпоративную систему хранения формируется квитанция, содержащая номер (ссылку) #INFOik в системе хранения, которая доставляется отправителю Xi, а #INFOik сообщается получателю Xj.

Ключ Kxi никому не известен и всегда находится внутри модуля доверенного хранения пользовательских ключей.

При запросе пользователя Xj на прочтение информации E(Kxi,INFOik) выполняется расшифрование E(Kxi,INFOik) и зашифрование INFOik на ключе Ksj внутри МДХКП, после чего результат передается пользователю Xj, который расшифровывает его на своем мобильном устройстве при помощи Ksj.

При хранении всех трех единиц – E(Kxi,INFOik), E(Ksi,INFOik) и E(Ksj,INFOik) –система получает полноценные свойства электронной подписи, с возможностью проверки в МДХПК и выдачи обоим пользователям результата верификации данных.

Если же участники системы не имеют возможности использовать модуль доверенного хранения пользовательских ключей, либо шифрование данных нецелесообразно, либо использование алгоритмов шифрования связано с экспортными или иными ограничениями, то ранее зарегистрировавшийся в системе участник дает запрос на проверку подписанного электронной подписью массива INFO тому участнику системы, который имеет открытый ключ или сер-

тификат для проверки. Этот участник проверяет ЭП и сообщает результат в виде отдельного массива, который мы назовем *CHECK* (содержащего в обязательном порядке и проверяемый документ) в распределенный реестр, оператор которого при положительном результате проверки ЭП связывает для других участников запись *CHECK* с первичным запросом INFO и тем самым дает возможность другим участникам системы проверить документ сравнивая содержания INFO и *CHECK*.

СМАРТ-КОНТРАКТ И ОРАКУЛЫ

Другим вариантом взаимодействия в системе документооборота может служить следующая схема.

Стороны документооборота договариваются и составляют смарт-контракт, в котором прописывают необходимые условия (в том числе порядок выбора посредника в случае необходимости). Например, что в определенный промежуток времени должен поступить перевод с некоторого адреса на соответствующий аккаунт смарт-контракта на определенную сумму и что в определенный промежуток времени в учетной системе РР должны появиться соответствующие блоки.

Если все условия выполнены, то средства будут переведены, если нет, то в течение определенного периода времени они возвращаются на исходный адрес. При таком механизме расчета минимизируются риски мошенничества любой из сторон и исключается необходимость доверия.

Важно отметить, что смарт-контракт может оперировать только теми данными, создание, хранение и обработка которых осуществляется в пределах одной учетной системы РР. Это означает, что валидаторы должны иметь доступ к данным, которые используются в смарт-контракте (такие как токены, балансы, транзакции, временные метки и др.). Если средства, участвующие в переводе, не учитываются в блокчейне, то необходимо использовать оракулы, т. е. такие доверенные третьи стороны, которые получают информацию от *offchain*-ресурсов и поставляют ее в блокчейн. Здесь возникает вопрос о достоверности и полноте этих данных. В настоящее время самыми распространенными подходами к решению этой проблемы являются консенсус оракулов (примером может служить проект Chainlink) или использование TLSNotary-доказательств для подтверждения корректной работы оракула. Необходимо подчеркнуть, что оракул является именно поставщиком данных, а не их источником. Поэтому, несмотря на то, что оба подхода в некоторой степени гарантируют корректность передачи данных от источника к смарт-контракту, они не могут обеспечить достоверность данных самого источника. Как следствие, рекомендуется обращаться к нескольким доверенным источникам данных.

Использование оракулов требует доверия пользователей к передаваемым данным. Эти данные должны быть подписаны электронной подписью оракула, а пользователи должны иметь возможность убедиться – какие данные и какой оракул передал на вход смарт-контракту.

Таким образом, с использованием оракулов также нельзя полностью исключить влияния третьей стороны. Оракул является потенциальной точкой отказа, так как существуют риски предоставления ошибочных данных при сбое, злонамеренном действии, компрометации или отказе в обслуживании. Для минимизации уровня доверия рекомендуется использовать нескольких оракулов, а также предусмотреть возможность выбора и обновления списка оракулов и их мотивацию при разработке смарт-контракта. Имеет смысл продумать возможность обновления смарт-контракта (в случае некорректной работы) еще на этапе его проектирования, если такая функциональность не реализована на платформе смарт-контрактов (как например в EOS).

ИСПОЛЬЗОВАНИЕ ДОКАЗАТЕЛЬСТВ С НУЛЕВЫМ РАЗГЛАШЕНИЕМ

Отметим, что в общем случае достаточно сложно обеспечить конфиденциальность входных данных и используемых в смарт-контракте условий.

Одним из наиболее перспективных направлений для обеспечения приватности пользователей считаются доказательства с нулевым разглашением (*zk-proofs*). Существуют интерактивные доказательства, состоящие из последовательности действий запросов и ответов между доказывающим и проверяющим (*commitment-challenge-response*), и неинтерактивные (*Random oracle model*, *Fiat-Shamir heuristic*, *Pairing-based* доказательства и др.), когда от проверяющего не требуется непосредственного опроса доказывающего. Особый интерес для децентрализованных учетных систем представляют именно неинтерактивные доказательства ввиду отсутствия необходимости дополнительного взаимодействия между участниками.

Необходимо заметить, что в зависимости от модели учета внутри системы (UTXO или аккаунты/балансы) для обеспечения валидности транзакций могут быть разные правила. Таким образом, требования и архитектура доказательств с нулевым разглашением систем для реализации транзакций с сохранением конфиденциальности отличаются для каждой модели [4].

В контексте нашей задачи можно использовать zk-SNARKs – неинтерактивные доказательства с нулевым разглашением (Non-Interactive Zero Knowledge protocol, NIZKP), имеющие некоторые ограничения:

- требования к вычислительным ресурсам и размер самих доказательств зависят от конкретного протокола, реализации и математической модели, на основе которых работают данные доказательства. Обычно транзакции, использующие доказательства с нулевым разглашением, имеют гораздо больший размер, что накладывает определенные требования на устройства конечных пользователей;
- zk-SNARKs являются компактными доказательствами, и к их основному недостатку можно отнести требование к наличию процедуры доверенной установки (*trusted setup*) для начальной настройки (при взаимодействии нескольких участников). На этом этапе происходит генерация открытых общесистемных параметров, использование которых позволяет производить проверку транзакций на соответствие правилам протокола. В данном случае верификаторы генерируют специальный секрет, ко-

торый должен быть сразу же уничтожен после доверенной установки, так как его существование допускает публикацию ложных доказательств (*fake proofs*). Однако в реальности не существует способа проверить факт удаления секрета. Эти риски можно снизить, выполняя установку секрета с использованием протокола конфиденциального вычисления (*multiparty computation – MPC*). В этом случае достаточно, чтобы хотя бы один верификатор был честным и удалил свой *toxic waste* после процедуры.

Из-за этого ограничения SNARK-протоколы плохо подходят для произвольных Тьюринг-полных смарт-контрактов, так как каждый новый контракт потребует новой установки. Интересным предложением по имплементации приватных смарт-контрактов может служить система Hawk [5], однако она не только требует новой установки на каждый контракт, но и необходимость использования так называемого «доверенного менеджера», который имеет доступ к приватным данным пользователя. Другой альтернативой реализации приватных смарт-контрактов могут служить СК типа Bulletproofs, так как здесь не требуется проведения процедуры доверенной установки и есть возможность реализовать обобщенный ZKP-протокол с относительно небольшими по размеру доказательствами [6, 7]. Существуют и модификации zk-SNARK [8], не требующие процедуры доверенной установки.

ЗАКЛЮЧЕНИЕ

Предлагаемая в настоящей статье технология создания платформы доверенного документооборота с электронной подписью может стать основой для создания международной системы доверенного документооборота с электронной подписью, предназначенной для взаимной проверки и признания документов, сформированных в различных юрисдикциях и снабженных электронной подписью, а соответственно, и для выполнения юридически значимых действий в рамках обмена электронными документами [9]. За техническую основу платформы может быть взят способ надежного доверенного хранения, обмена и проверки документов на основе доверенного распределенного реестра, с использованием смарт-контрактов и симметричных криптографических алгоритмов.

СПИСОК ЛИТЕРАТУРЫ

1. Электронная подпись: трансграничное взаимодействие. – URL: https://habr.com/ru/post/144401/
2. Трансграничный электронный документооборот. – URL: https://ca.gisca.ru/solutions/transgranichnyy-elektronnyy-dokumentooborot/
3. Кузьменко В.В., Макаров В.Л., Разгуляев К.А., Хан Д.В., Щербаков А.Ю. Новый подход к обеспечению безопасности периметра бизнес-процессов и аутентификации пользователей в корпоративной системе // Вестник современных цифровых технологий. – 2020. – №3. – С.10-13.
4. ZKProof. ZKProof Community Reference. Version 0.2 / eds. by D. Benarroch, L.T.A.N. Brandão, E. Tromer. Pub. by zkproof.org. Dec. 2019. – URL: https://zkproof.org
5. Kosba Ahmed, Miller Andrew, Shi Elaine, Wen Zikai. Charalampos Papamanthou. Hawk: The Blockchain Model of Cryptography and Privacy-Preserving Smart Contracts. – URL: https://eprint.iacr.org/2015/675.pdf (дата обращения: 28.04.2020).
6. Bünz B., Agrawal S., Zamani M., Boneh D. Zether: Towards privacy in a smart contract world. – URL: https://crypto.stanford.edu/~buenz/papers/zether.pdf (дата обращения: 23.04.2020).
7. Lin Dr. Suterusu yellowpaper. – URL: https://github.com/suterusu-team/Suter_yellowpaper/blob/master/Suterusu%20yellowpaper%20V%200.2.pdf (дата обращения: 25.04.2020).
8. Bünz Benedikt, Fisch Ben, Szepieniec Alan. Transparent SNARKs from DARK Compilers. – URL: https://eprint.iacr.org/2019/1229.pdf (дата обращения: 20.04.2020).
9. Шаститко А.Е., Шпакова А.А. Регулирование трансграничного электронного документооборота в евразийском экономическом союзе // Государственное управление. Электронный вестник. – 2018 г. – URL: https://cyberleninka.ru/article/n/shastitko-a-e-shpakova-a-a-regulirovanie-transgranichnogo-elektronnogo-dokumentooborota-v-evraziyskom-ekonomicheskom-soyuze/viewer.

Материал поступил в редакцию 16.10.20.

Сведения об авторе

ЩЕРБАКОВ Андрей Юрьевич – доктор технических наук, профессор, руководитель Центра развития криптовалют и цифровых финансовых активов ВИНИТИ РАН, профессор кафедр «Интеллектуальные системы информационной безопасности» технического университета МИРЭА и «Безопасность цифровой экономики и управления рисками» Российского государственного университета нефти и газа имени И.М. Губкина, консультант по криптографическим технологиям фирмы «Лаборатория Касперского», Москва
e-mail: x509@ras.ru

УДК [002:004]–047.44

В.В. Миронов

Обработка данных о публикационной активности автора в составе авторского коллектива с учетом квартилей журналов

Рассматриваются результаты краткого исследования информативности известного и простого в применении наукометрического индекса цитирования Хирша (оценки публикационной активности автора), показаны его достоинства и недостатки. Представлены как альтернатива два новых индекса цитирования: 1) трехмерный A-индекс, вычисляемый с учетом вклада авторского коллектива и квартилей журналов; 2) одномерный N-индекс, равный псевдонорме A-индекса. Для удобства сравнения публикационной активности автора на множестве всех (учитываемых) авторов введены соответствующие отношения порядка. Все предлагаемые индексы просты в использовании, включают, как часть, привычную для научного сообщества методику Хирша и одновременно «на порядок» информативнее последнего.

Ключевые слова: *обработка данных, публикационная активность, индекс Хирша, новые информативные индексы цитирования, эффективность оценивания*

DOI: 10.36535/0548-0027-2020-11-4

ИНДЕКС ХИРША

Наукометрический числовой показатель цитирования индекс Хирша (или *h*-индекс) предложен в 2005 г. американским физиком аргентинского происхождения Хорхе Хиршем (J.E. Hirsch) из Калифорнийского университета (Сан Диего, США) для оценки публикационной активности ученых-физиков [1]. Этот индекс выражает количественную характеристику продуктивности учёного (или группы учёных, или научной организации, или даже страны в целом) и исчисляется на основе количества публикаций и количества цитирования (в любых изданиях) публикаций этого ученого. Из-за простоты вычисления и необходимости хоть как-то количественно оценивать работу ученых со стороны чиновников, индекс Хирша распространился на публикационную деятельность всех научных направлений.

Излишнее внимание к количественной стороне научного творчества порождает определенные проблемы, в частности, стимулирует неоправданное самоцитирование, тиражирование публикаций (пересказ прежних статей без новых результатов – самоплагиат), использование подчиненных сотрудников и учеников для искусственного повышения индекса цитирования ученого.

Индекс Хирша вычисляется просто: если h статей ученого из их общего числа N_p цитируются h (или более) раз каждая, и каждая из оставшихся N_p - h статей цитируется менее (или ровно) h раз, то h-индекс ученого равен (натуральному) числу h (если же нет статей или ссылок на них, то h=0).

Считается, что с помощью индекса Хирша оценивается «ядро цитирования» (в интуитивном его понимании) или «ядро публикационной активности ученого». Он формируется через Интернет на основе свободных в доступе наукометрических баз публикационных данных (например, Google Scholar, Elibrary.ru, ADS NASA), а также на основе платных баз данных (Scopus, Web of Science и др.) [2–6].

Индекс Хирша не оптимален для своих же целей, его недостатки отмечены в основополагающей статье самого ученого [1]. Главный недостаток этого показателя в том, что соотношение h-индексов ученых зачастую не соответствует их вкладу в развитие соответствующей отрасли науки (подразумевается, что чем больше h-индекс, тем больше вклад или «вес» ученого в развитии его отрасли науки).

Например, если бы h-индекс существовал во времена Э. Галуа, то его h-индекс был бы равен 4, а h-индекс А. Эйнштейна в 1906 г. – всего лишь 5, несмотря на очень высокий показатель цитирования его произведений в 1905 г. [7] и одновременно значительный вклад обоих гениев в развитие науки! Известны и обратные примеры, но… «об отсутствующих или хорошо или ничего».

Положительные стороны h-индекса также очевидны: простота вычисления, доступность данных и «в первом приближении» неплохая оценка публикационной активности ученого. При этом имеется в виду, что, как говорил Британский Премьер У. Черчилль, «демократия – наихудшая форма правления, если не считать всех остальных». (Заметим, что этот известный афоризм красив, но внутренне противоречив. В са-

мом деле, отношение «наихудшая» – это бинарное отношение и «не считать всех остальных» просто нельзя!)

Недостатки количественной оценки вклада ученого в науку, предложенной Х. Хиршем, породили *многочисленные* предложения по её улучшению и на основе методики Хирша, и на других принципах (см., например, [8–18], и особо работу Л. Вальтмана и Н.Й. ван Экка [19]. Еще почти четыре десятка работ с различными предложениями по модернизации идеи Хирша или «удаленными» от неё нововведениями в библиометрию можно найти на сайте [20]).

В настоящей работе предлагаются два новые индекса: 1) трехмерный *A*-индекс, вычисляемый с учетом вклада авторского коллектива и квартилей журналов и 2) одномерный *N*-индекс, равный (по определению) некоторой псевдонорме *A*-индекса. Новизна этой идеи (по сравнению с идеей Х. Хирша) достигается за счет введения в рассмотрение новых понятий – так называемых «хвоста» и «подвала» цитирования.

Оба предлагаемых нами индекса, сохраняя идею Хирша о «ядре цитирования» работ ученого, а также простоту и наглядность оценки, на порядок превосходят классический *h*-индекс по *качеству* оценки вклада (или влияния) ученого в развитие соответствующей отрасли науки на основе данных о цитирования его работ (корреляция «чем больше публикаций, тем больше вклад ученого» сомнению здесь не подвергается).

Одновременно, для удобства сравнения публикационной активности автора, на множестве всех (учитываемых) авторов мы вводим соответствующие отношения порядка, с помощью которых строим *количественную оценку* вклада ученого в развиваемую им отрасль, по значению этой оценки (кому-то) удобно «ранжировать» представителей науки.

Эти качества – преемственность простоты вычисления и «прозрачности» индекса Хирша при одновременной на порядок более высокой объективности оценки вклада ученого в его отрасль науки – выгодно отличают предлагаемые нами процедуры от упомянутых в работах [8–20].

A-ИНДЕКС ЦИТИРОВАНИЯ

Введем необходимые определения, данные для алгоритмов, и дадим некоторые комментарии.

Определение 1. Расположим все статьи автора *NN* списком сверху вниз по мере убывания числа ссылок на его статьи (цитирования) в журналах, включенных в Перечень ВАК, и в более «высоких» базах данных – Scopus, Web of Science. Статьи с одинаковым числом цитирований расположим в любом порядке. Получим список под условным названием «*P-список*» публикаций автора *NN* с общим числом работ N_p (по определению $N_p > 0$).

Представим *P-список* в виде таблицы или (квадратной) матрицы (которую назовем *P*-таблицей или *P*-матрицей). По построению у *P*-матрицы «по вертикали» располагаются все статьи (удобнее, если статьи просто пронумеровать) автора *NN*, вошедшие в *P-список*, а по «горизонтали» в каждой строке, соответствующей выбранной статье, стоит отметка о цитировании (такие отметки также удобно пронумеровать) соответствующей статьи каким-то автором в каком-то журнале (более конкретной информации, кто именно и где именно, не требуется).

Размерность такой *P*-матрицы есть $N_p \times M_p$, где число работ N_p определено ранее, а M_p – число цитирований первой статьи автора в журнале, который входит в одну из учитываемых в данном случае баз цитирования.

Ячейку (элемент *P*-матрицы) на пересечении вертикальной (о статьях) и горизонтальной (о цитировании) информации назовем непустой, а другие ячейки, для которых подобная информация отсутствует, – пустыми.

Непустую ячейку (элемент) матрицы заполним числом 1 (единицей), пустую ячейку заполним числом 0 (нулем).

С каждой цитируемой статьей (стоящей на i-ом месте в *P*-списке) свяжем пару чисел $p_i = (n_i, Q_i)$, где n_i – число авторов i-й статьи, Q_i – квартиль журнала, в котором была опубликована статья.

Журналам из международных баз данных присваивают, как известно, один из четырех квартилей: $Q_1 = 1$, $Q_2 = 2$, $Q_3 = 3$, $Q_4 = 4$. Если журнал из какой-то базы (например, из Перечня ВАК) не имеет квартиля, то присвоим ему квартиль Q_4.

По **Определению 2 *«ядро цитирования»*** *A*-индекса – это квадратная таблица размером $h \times h$, где сторона h соответствует (численно равна) индексу Хирша. При этом «физические» размерности (или измерения) сторон квадрата разные: первое h – это число статей, второе h – минимальное число цитирований каждой из статей, и хотя бы одна статья имеет ровно h цитирований.

По **Определению 3 *«рубеж цитирования»*** – это (достаточно большое) число Δ (имя – по определению), являющееся предельным для «полновесного» учета числа цитирований той или иной статьи. Если число цитирований статьи превысило число Δ, то эти превышающие число Δ цитирования входят в зачет *A*-индекса с коэффициентом $k < 1$. В настоящей рабате, в ее алгоритмической части, мы предлагаем $k = 0{,}5$.

Для чего вводится «рубеж цитирования»? Дело в том (так уж сложилось в научном сообществе), что, начиная с некоторого «предельного» числа цитирований возникает своего рода «мода» на автора, на цитирование его статей по соответствующей теме (или по соответствующей статье).

Ученые (в особенности молодые) считают престижным ссылаться именно на известного, «модного» автора, придавая тем самым определенный «вес» своим работам. Такая «мода» в свое время была, например, на цитирование работ (произведений) В.И. Ульянова (Ленина), Ж.А. Пуанкаре, А.Н. Колмогорова, А.С. Пушкина.

Сегодня наиболее цитируемые авторы – это относительно небольшая группа лауреатов Нобелевской премии, а также группа быстро прогрессирующих специалистов в разных областях науки; их

имена, списки их публикаций можно легко найти в Интернете (сохраняя при этом определенный скептицизм по отношению к разным классификационным спискам).

Анализируя ситуацию, в том числе принимая во внимание метазакон Мерфи – «чем больше ты на публике, тем больше тебя приглашают», – можно считать обоснованным введение такого предельного числа Δ в определении и продвижении *A*-индекса цитирования.

Исходя из анализа открытых индексов цитирования ученых, мы предлагаем положить «рубеж цитирования» состоящим из двух частей:

часть *A*. Если индекс $h \le 50$, то вводится (фиксированный) рубеж цитирования $\Delta = 100$;

часть *B*. Если индекс $h > 50$, то вводится (плавающий) рубеж цитирования $\Delta = 2h$.

ЗАМЕЧАНИЕ. В части *A* фигурирует индекс $h \le 50$, для которого вводится (фиксированный) рубеж цитирования $\Delta = 100$. Заметим, что индекс Хирша в 50 единиц – это очень высокий показатель. Достаточно посмотреть на индексы Хирша (размещенные в Интернете) российских ученых – членов РАН или членов Американской академии наук.

И одновременно без *h*-индекса более 50 трудно рассчитывать на кафедру, к примеру, в MIT.

По **Определению 4 «*хвост списка цитирования*»** – это часть *P*-таблицы, у которой «по вертикали» расположены только статьи (или их номера), названия которых входят в ядро цитирования, а по «горизонтали» в каждой строке, соответствующей выбранной статье, –числа цитирований автора в журналах без учета цитирований, уже вошедших в ядро, т.е. единицы в соответствующих ячейках матрицы.

По **Определению 5 «*подвал списка цитирования*»** – это таблица, у которой «по вертикали» расположены все статьи (или их номера), названия которых не вошли в ядро цитирования, но при этом каждая такая статья имеет ровно h цитирований, а «по горизонтали» – все цитирования этих упомянутых статей (h штук единиц).

Итак, пусть «ядро» цитирования, квадратная таблица размером $h \times h$, сформировано.

По **Определению 6** положим (знак умножения – точка, и множитель 1 сохранены в формулах (1) – (3) для понимания «физического смысла» вводимых определений):

$$S_0 = \sum_{i=1}^{h} (5 - Q_i) \cdot {}^{1}\!/_{n_i} \cdot 1 \cdot h; \qquad (1)$$

$$S_1 = \sum_{i=1}^{h} (5 - Q_i) \cdot {}^{1}\!/_{n_i} \cdot 1 \cdot (H_i - h) + \\ + k(5 - Q_i) \cdot {}^{1}\!/_{n_i} \cdot 1 \cdot (H_i - \Delta), \qquad (2)$$

где H_i –число цитирований i-й статьи в «хвосте» (или число единиц в соответствующей i-й строке *P*-матрицы), при этом если для какого-то i значение $H_i \le \Delta$, то для этого i в формулу (2) входит только одно левое слагаемое, а если значение $H_i > \Delta$, то для этого i в формулу (2) входят оба слагаемых;

$$S_2 = \sum_{i=1}^{r} (5 - Q_i) \cdot {}^{1}\!/_{n_i} \cdot 1 \cdot h, \qquad (3)$$

где r – число статей в «подвале» списка цитирования.

ЗАМЕЧАНИЕ. Прокомментируем формулы (определения) (1) – (3). Единица в формулах – это собственно одно цитирование *i*-й статьи, входящей либо в «ядро», либо в «хвост», либо в «подвал» списка цитирования. Множитель справа от 1 – это число всех цитирований *i*-й статьи, входящих либо в «ядро» (формула (1)), либо в «хвост» (формула (2)), либо в «подвал» (формула (3)). Коэффициент *k* (нами рекомендован *k* = 0,5) в формуле (2) снижает «долю цитирования» в $1/k$ раз.

Множитель ${}^{1}\!/_{n_i}$ – это вклад автора в работу коллектива (напомним, коллектив авторов *i*-й статьи состоит из n_i человек, вклад каждого автора считается равным, по определению).

Заметим также, что если вклад авторов в написание статьи неодинаков (что бывает нередко) и авторы хотят разделить этот вклад неравномерно (что бывает редко), то эта ситуация обрабатывается следующим образом: в формулу (3) вместо множителя ${}^{1}\!/_{n_i}$ вносится множитель $k = k_j$, где j – номер автора в коллективе из n_i человек с номерами $1, 2, \ldots, n_i$, соответственно, при этом сумма коэффициентов вклада $k_1 + k_2 + \ldots + k_{n_i} = 1$.

Множитель $(5 - Q_i)$ – это коэффициент увеличения «доли цитирования» за счет более высоких квартилей журналов, в которых была напечатана *i*-я статья.

Возвращаясь к продуктивной (по крайней мере, для многочисленных администраторов науки) идее «квадрата Хирша» размером $h \times h$ или «площадью» h^2, формула (1) определяет «площадь» взвешенного аналога квадрата Хирша, а формулы (2) и (3) – это аналоги «площадей» взвешенных квадратов, но для «хвоста» и «подвала» списка цитирования, соответственно.

После такого разъяснения следующее определение выглядит естественным.

По **Определению 7** *A*-индекс (или индекс *A*) публикационной активности ученого, рассчитанной на основе анализа цитирований его научных публикаций с учетом вклада авторского коллектива и квартиля журнал, где была опубликована статья, есть упорядоченная тройка (натуральных) чисел

$$A = (a, p, q), \qquad (4)$$

где компоненты определяются (и вычисляются) по следующим правилам:

$$a = \left[\sqrt{S_0}\right], \quad (5)$$

$$p = \left[\sqrt{S_1}\right], \quad (6)$$

$$q = \left[\sqrt{S_2}\right], \quad (7)$$

символ $[\ldots]$ в правых частях формул (5) – (7) – это обозначение стандартной функции целой части числа, заключенной в этих квадратных скобках, например, $[7,7]=7$.

A-индекс позволяет дать количественную и качественную взвешенную оценку ядра списка цитирования и его «окружения» в виде «хвоста» и «подвала», но тройка чисел (4) психологически трудно воспринимается (неподготовленным пользователем).

В связи с этим для удобства сравнения публикационной активности ученых (обозначим её $A(NN)$ для ученого *NN*), например, с целью поддержания финансирования работ или кадрового роста персонала, введем одномерный псевдонормированный индекс.

Определение 8. Положим, что одномерный псевдонормированный индекс

$$\delta = [M] = \left[\sqrt{a^2 + p^2 + q^2}\right], \quad (8)$$

в формуле (8) символ из двух квадратных скобок $[\ldots]$ – снова, как и в формулах (5) – (7), стандартная функция целой части числа.

Тогда следующее определение становится «прозрачным».

Определение 9. Для двух ученых NN_1 и NN_2 с *A*-индексами их публикационной активности $A_1 = (a_1, p_1, q_1)$ и $A_2 = (a_2, p_2, q_2)$ и псевдонормированными индексами δ_1 и δ_2, соответственно, положим:

$$A(NN_1) > A(NN_2) \overset{def}{\Leftrightarrow} \overset{def}{\Leftrightarrow} \begin{cases} \delta_1 > \delta_2, \\ \delta_1 = \delta_2 \wedge h_1 > h_2, \\ \delta_1 = \delta_2 \wedge h_1 = h_2 \wedge p_1 > p_2, \\ \delta_1 = \delta_2 \wedge h_1 = h_2 \wedge p_1 = p_2 \wedge q_1 > q_2. \end{cases} \quad (9)$$

(В формуле (9) символ конъюнкции $\wedge$ – это обозначение союза «и», связывающего соседние высказывания более слабо.

В противном случае, публикационные активности авторов считаются одинаковыми: $A(NN_1) = A(NN_2)$.

ЗАМЕЧАНИЕ. С целью детализации информации индексы a, p, q, δ можно вычислять с точностью до десятых долей.

АЛГОРИТМ ВЫЧИСЛЕНИЯ *A*-ИНДЕКСА

Приведем алгоритм вычисления *A*-индекса, давая при необходимости пояснения и осознавая одновременно его простоту. (Мы не вдаемся здесь в формальную строну дела, определяя различные виды сложности, а, значит, и простоты: битовую, арифметическую, временную и др.; по поводу «сложностей» можно посмотреть, например, [21]). Простота алгоритма здесь видна.

1. Начало алгоритма. Представим все статьи автора *NN* в виде *P*-таблицы.

2. По *P*-таблице формируем *P*-матрицу необходимой размерности и заполняем ее ячейки надлежащим образом единицами и нулями.

3. В *P*-матрице из общего числа N_p выделяем h статей, на каждую из которых приходится h или более цитирований, а на каждую из оставшихся $N_p - h$ статей приходится h или менее цитирований.

Следовательно, индекс Хирша автора *NN* равен h. И, следовательно, сформировано ядро *A*-индекса – квадрат $h \times h$.

4. Формируем «хвост» и «подвал» списка цитирования.

5. Вычисляем параметры a, p, q по формулам (5) – (7).

6. Формируем *A*-индекс автора *NN* по формуле (4).

7. Формируем δ-индекс автора *NN* по формуле (8).

8. Конец алгоритма.

НЕКОТОРЫЕ СВОЙСТВА НОВЫХ ИНДЕКСОВ

Представленные далее утверждения являются простыми следствиями определений, однако их явная формулировка позволяет быстрее оценить некоторые свойства предложенного нами метода.

***Утверждение* 1**. Новая методика сохраняет преемственность методики Хирша.

Доказательство. В самом деле, достаточно взглянуть на определения, данные в виде формул (4) – (7), и проследить «физическую сущность» этих определений в виде сторон взвешенных квадратов ядра индекса Хирша и сопутствующих квадратов «хвоста» и «подвала» списка (что один для одного автора) цитирования.

***Утверждение* 2.** Если все учитываемые в нашем исследовании статьи написаны одним автором, опубликованы в журналах 4-го квартиля, то для такого автора $a = h$, где h – индекс Хирша. Для такого случая $\delta \geq h$.

***Утверждение* 3**. В подвале «ядра» списка цитирования не может быть статей с числом цитирований больше, чем *h*.

Доказательство. В самом деле, в противном случае либо эта статья была бы уже учтена в ядре цитирования, либо «ядро» имело бы размеры $s \times s$, при $s > h$, что противоречит его определению.

Определение 10. Будем считать, что индекс X (где синтаксическая переменная X может принимать значения $X = a, p, q, A, \delta$) устойчив в области D (где синтаксическая переменная D может принимать значения D = «ядро», «хвост», «подвал»), если никакие одно-два-три «случайных» цитирования (но-

вых или старых) статей автора из области D не может с большой долей вероятности изменить значение индекса X.

В противном случае индекс X неустойчив в области D.

Будем считать, что индекс X глобально устойчив, если он устойчив при всех значениях переменной D.

Тезис 1. Индексы a, p, q – устойчивые характеристики в своих областях; индекс δ – глобально устойчивая характеристика, более устойчивая, нежели индекс Хирша.

Доказательство. В самом деле, нетрудно посчитать классические (статистические) вероятности изменения характеристик a, p, q, δ при одном и двух «случайных» цитированиях (новых или старых) статей автора из области D и убедиться в относительно небольших значениях этих вероятностей.

Из определения 8 (что дано в виде формулы (8)) и проведенных (гипотетических) подсчетов вытекает, что индекс δ на порядок устойчивее (при интуитивном понимании «порядка») индекса Хирша.

ЗАМЕЧАНИЕ. Читатели (они же авторы), без сомнения, отметили свойство устойчивости индекса Хирша на своем личном опыте.

Тезис 2. Индекс A на порядок информативнее индекса Хирша.

Этот тезис (равно, как и заключительную часть тезиса 1) нельзя доказать строго. Для его математического доказательства нужно дать строгие определения «информативности» и «объективности», ввести и проранжировать их порядки. Это, в любом случае, спорная тема, и потому здесь не рассматривается.

Одновременно можно не строго, а на содержательном уровне обосновать этот тезис, основываясь на принципах интуиционистской математики (основы которой можно почерпнуть, например, в работах [22–24]). Вспомним здесь и высказывание Д. Пойа: «Конечно, будем учиться доказывать, но будем также учиться догадываться» [25].

Нельзя не согласиться, что информация, вошедшая в «хвост» и «подвал» ядра списка цитирования, важна для общей оценки публикационной активности автора и, как следствие, его (автора) кадровых, должностных перспектив (конечно, при наличии других, не менее важных характеристик). Одновременно, если не учитывать такую информацию, то существенно сужается общее представление об ученом, точнее, о количестве и качестве его публикаций.

В таких оценках более «мелкой» информацией, не вошедшей в «хвост» и «подвал» списка цитирования, можно пренебречь.

ПРИМЕРЫ

Проиллюстрируем обоснованность претензий A-индекса на большую объективность в оценке публикационной активности автора по сравнению с индексом Хирша при сопоставимой простоте вычисления.

ПРИМЕР 1. Автор NN_1 опубликовал 30 работ. Каждая из них написана им одним; 10 статей опубликованы в журналах 3-го квартиля и имеют по 10 цитирований каждая; остальные 20 работ опубликованы в журналах 4-го квартиля и имеют по 5 цитирований.

Каков индекс публикационной активности автора NN_1?

Нетрудно видеть, что индекс Хирша ученого NN_1 есть число $h_1 = 10$. Посчитаем активность ученого по предложенной в настоящей статье методике. Получим $A = (14, 0, 0)$, $\delta = 14$. Таким образом, индекс Хирша автора увеличился за счет «взвешенного» анализа, но общий вклад автора в его отрасль остался без изменений.

ПРИМЕР 2. Автор NN_2 опубликовал 30 работ, из них 10 статей – в журналах 3-го квартиля, написаны в соавторстве с еще двумя авторами и имеют по 100 цитирований каждая; остальные 20 работ написаны в соавторстве с одним коллегой, опубликованы в журналах 4-го квартиля и имеют по 10 цитирований.

Каков индекс публикационной активности автора NN_2?

Нетрудно видеть, что индекс Хирша ученого NN_2 есть $h_2 = 10$.

Посчитаем активность ученого по предложенной в настоящей статье методике. Получим что $A = (8, 24, 10)$, $\delta = 27$. Таким образом, индекс Хирша автора уменьшился за счет «взвешенного» анализа, но общий его вклад в отрасль науки увеличился почти в 3,5 раза!

ЗАКЛЮЧЕНИЕ

Настоящая работа – попытка отойти от принятых шаблонов и схем в оценке деятельности научных и педагогических работников и, особенно, в оценке перспектив ученого или педагога. В этом контексте настоящая статья примыкает к публикации [26].

По нашему мнению, учет всей совокупности информации, вошедшей в «ядро», «хвост» и «подвал» ядра списка цитирования, существенно расширяет общее представление о публикационной активности ученого, как в количественном, так и в качественном ее аспектах, по сравнению с методикой Хирша. Если такая информация не учитывается, то это существенно сужает общее представление о публикационной активности ученого, как в количественном, так и в качественном ее аспектах. Другой информацией, как видится, можно пренебречь.

Однако, бесспорно, эта информация будет более точной, если её скорректировать, как предложено в настоящей работе, по каждой статье с учетом научного вклада и коллектива авторов статьи, и квартиля журнала, где она была опубликована.

* * *

Автор выражает благодарность Кристине Валентиновне Мироновой, кандидату технических наук, ведущему специалисту крупной IT-компании, за полезные советы при написании настоящей работы, за просмотр и анализ обширной интернет-информации по интересовавшему вопросу, а также редактирование текста статьи.

СПИСОК ЛИТЕРАТУРЫ

1. Hirsch J.E. An index to quantify an individual's scientific research output // Proceedings of the National Academy of Sciences of the United States of America. – 2005. – Vol. 102, № 46. – P. 16569-16572.
2. Google Scholar. – 2020. – URL: https://scholar.google.com/ (дата обращения: 21.06.2020)
3. Elibrary.Ru, Научная Электронная Библиотека. – 2020. – URL: https://elibrary.ru/ (дата обращения: 21.06.2020)
4. Astrophysics data system (ADS NASA). – 2020. – URL: http://adsabs.harvard.edu/ (дата обращения: 22.06.2020)
5. Bar-Ilan J. Which h-index? – A comparison of WoS, Scopus and Google Scholar // Scientometrics. – 2007. – Vol. 74, №. 2. – P. 257—271.
6. Рейтинг науковців України за показниками наукометричної бази даних Scopus. 05.12.2013, Архивная копия от 5 октября 2013 на Wayback Machine. – 2013. – URL: https://webarchiveorg.ru/ (дата обращения: 26.06.2020)
7. Михайлов О.В., Михайлова Т.И. Индекс Хирша в оценке деятельности ученого в национальном исследовательском университете // Вестник Казанского технологического университета. – 2010. – № 11. – С. 485-487.
8. Tagiew R., Ignatov D.I. Behavior mining in h-index ranking game // CEUR Workshop Proceedings. – 2017. – Vol. 1968. – P. 52—61.
9. Имаев В. Технологии увеличения индекса Хирша и развитие имитационной науки // Комиссия РАН по борьбе с лженаукой и фальсификацией научных исследований. В защиту науки. – 2016. – № 17. – С. 38—51.
10. Демина Н. Хиршемания и хиршефобия. «Троицкий вариант – Наука». – 2016. – URL: https://trv-science.ru/2016/12/06/khirshemaniya-i-khirshefobiya/ (дата обращения: 25.06.2020).
11. Михайлов О.В. Новая версия индекса Хирша – *j*-индекс // Вестник РАН. – 2014. – Т. 84, № 6. – С. 532-535.
12. Egghe L. Theory and practise of the g-index // Scientometrics. – 2006. – Vol. 69, № 1. – P. 131-152.
13. Kosmulski M.I. A bibliometric index // Forum Akademickie. – 2006. – Vol. 11. – P. 31.
14. Prathap G. Hirsch-type indices for ranking institutions' scientific research output // Current Science journal. – 2006. – Vol. 91(11). – P. 1439.
15. Jones T., Huggett S., Kamalski J. Finding a Way Through the Scientific Literature: Indexes and Measures // World Neurosurgery. – 2011. – Vol. 76. - № 1, 2. – P. 36-38.
16. Холодов А.С. Об индексах цитирования научных работ // Вестник РАН. – 2015. – Т. 85, № 4. – С. 310-320
17. Мазов Н.А., Гуреев В.Н. Альтернативные подходы к оценке научных результатов // Вестник РАН. – 2015. – Т. 85, № 2. – С. 115-122.
18. Кузнецов А.В. Для начала надо навести порядок в существующей системе РИНЦ. Письма в редакцию // Вестник РАН. – 2014. – Т. 84, № 3. – С. 268-269.
19. Ludo Waltman, Nees Jan van Eck. Robust Evolutionary Algorithm Design for Socio-Economic Simulation: Some Comments // Comput. Econ – 2009. – Vol. 33. – P.103–105
20. hα: An index to quantify an individual's scientific leadership. – 2020. – URL: http://link.springer.com/article/10.1007/s11192-018-2994 (дата обращения: 26.06.2020)
21. Разборов А.А. О сложности вычислений // Математическое просвещение. – 1999. – № 3. – С. 127 -141.
22. Вейль Г. О философии математики. Сборник работ. – М.: ГИТТЛ, 1934. – 128 с.
23. Гейтинг А. Интуиционизм. – М.: Мир, 1965. – 202 с.
24. Френкель А.А., Бар-Хиллел И. Основания теории множеств. – М.: Мир, 1966. – 553 с.
25. Пойа Д. Как решать задачу. – М.: ГУ-ПИМП, 1959. – 205 с.
26. Миронов В.В. Информатизация образования: достижения и проблемы // Информатизация образования и науки. – 2017. – № 4(36). – С. 3-18.

Материал поступил в редакцию 25.07.20.

Сведения об авторе

МИРОНОВ Валентин Васильевич – доктор физико-математических наук, профессор кафедры высшей математики, директор лаборатории системного анализа, Рязанский государственный радиотехнический университет имени В. Ф. Уткина.
e-mail: mironov1vv @mail.ru

АВТОМАТИЗАЦИЯ ОБРАБОТКИ ТЕКСТА

УДК 81’322.2:54-3

Н.Ю. Бизюкова, О.А. Тарасова, А.В. Рудик, Д.А. Филимонов, В.В. Поройков

Автоматическое распознавание названий химических соединений в текстах научных публикаций*

Рассмотрены методы поиска и извлечения наименований низкомолекулярных химических соединений и данных об их экспериментально подтверждённой биологической активности из текстов научных публикаций. Проанализированы разработанные и опубликованные в течение последних десяти лет подходы для автоматизированного извлечения химической и биологической информации, представленной (а) наименованиями химических соединений и (б) наименованиями белков, генов и ассоциированных с ними видов биологической активности. Такие данные могут быть применены для идентификации и хранения названий химических соединений, включая все их возможные синонимы. Тематика научных публикаций весьма разнообразна, поэтому извлеченные данные о названиях химических соединений могут быть применены для получения информации о (1) способах синтеза определённого химического соединения; (2) его физико-химических свойствах; (3) его взаимодействии с высокомолекулярными соединениями (белками, мРНК животных и человека, и пр.) или проявлении им определённого вида биологической активности; (4) его терапевтических свойствах и данных клинических исследований.

Ключевые слова: *интеллектуальный анализ текстов, наименования химических соединений, информационный поиск*

DOI: 10.36535/0548-0027-2020-11-5

ВВЕДЕНИЕ

Процесс извлечения данных из слабо структурированных и формализованных текстов научных публикаций требует немалых усилий и затрат времени, особенно в условиях работы с большим объемом информации. В связи с этим возникает необходимость интеллектуального анализа текстов с помощью машинных методов их автоматизированной обработки, которые могут применяться в различных областях, включая медицину и биологию. Одно из направлений, требующих работы с большими массивами структурированных данных, – это биоинформатика. Полученные в результате анализа данные могут быть применены, в частности, для оценки биологической активности и токсичности химических соединений, а это необходимо при разработке новых лекарственных препаратов.

Методы извлечения данных из текстов рассмотрены в нескольких публикациях, включая обзор по методам обработки текстов биомедицинской тематики [1], а также описание и критический анализ методов извлечения данных из статей биологической и медицинской направленности [2–5].

В настоящей статье мы проанализируем методы извлечения названий химических соединений (ХС) из текстов научных публикаций, разработанные и опубликованные в течение последних десяти лет и рассмотрим методы автоматического извлечения данных о взаимодействии ХС с белками человека, приводящие к конкретным биологическим эффектам. В отличие от ранее опубликованных работ нами подробно рассматриваются корпуса (коллекции текстов), специально подготовленные для применения методов извлечения данных, а также способы представления текстов для их обработки компьютерными методами и алгоритмы извлечения данных о названиях химических соединений.

ИЗВЛЕЧЕНИЕ ДАННЫХ ИЗ ТЕКСТОВ

В интеллектуальном анализе текстов для извлечения данных о химических соединениях применяются как стандартные методы автоматизированной обработки текстов, так и специальные алгоритмы, направленные на поиск слов и словосочетаний, которые могут являться данными о ХС или их свойствах.

* Работа выполнена при поддержке гранта Российского научного фонда № 19-15-00396.

Рассмотрим представленные в электронном виде тексты, которые можно получить из библиографических баз данных (БД), например, БД Medline [6] в форматах PDF (Portable Document Format – межплатформенный открытый формат электронных документов, разработанный Adobe Systems Inc.), HTML (HyperText Markup Language – язык гипертекстовой разметки) и XML (eXtensible Markup Language — расширяемый язык разметки для автоматизированного создания и обработки документов).

Можно выделить следующие стандартные методы анализа текстов: предварительная обработка (препроцессинг), включающая (1) конвертацию наиболее распространённых форматов в простой текстовый формат; (2) разделение текста на элементарные единицы текста – токены, которые могут быть представлены как отдельными словами, так и различными символами и знаками препинания (токенизация); (3) классификацию токенов по принадлежности к частям речи (установку тегов соответствующих частей речи, Parts of Speech tags – PoS), приведение отдельных слов к словарной форме (лемматизация), удаление слов и терминов, которые часто встречаются, но не отражают смысловое содержание текстов узкоспециализированной, например, химической тематики (так называемые стоп-слова).

2-Acetyloxybenzoic acid

Рис. 1. Пример двух вариантов токенизации для ацетилсалициловой кислоты (наименование ИЮПАК 2-Acetyloxybenzoic acid).
Информация о вариантах наименований ацетилсалициловой кислоты получена из БД PubChem (https://pubchem.ncbi.nlm.nih.gov/).

Рис. 2. Принципы работы алгоритмов по извлечению информации из текстов и установлению ассоциаций между терминами.

Токенизация – это разделение текста на слова, числа или другие специальные обозначения, или выделение нескольких слов, потенциально относящихся к одному и тому же термину. Алгоритм токенизации тривиален для простого текста с очевидными разделителями слов (пробелами). На практике проблемы токенизации возникают, когда в границах слова есть дефисы, запятые, скобки, апострофы – все это довольно типичная ситуация для названий химических соединений (рис. 1). Алгоритмы токенизации, как правило, базируются на применении методов машинного обучения к большому массиву текстов, где разметка токенов выполнена вручную. Выбор алгоритма токенизации важен при формировании словарей синонимов названий ХС на основе текстов научных публикаций [7, 8].

Разработку методов извлечения из текстов данных о химических соединениях и определения их взаимодействия с биологическими объектами можно условно разделить на две самостоятельные обширные задачи: (1) извлечение данных о ХС и биологических объектах и (2) поиск ассоциаций между ними (рис. 2).

ИЗВЛЕЧЕНИЕ ДАННЫХ О ХИМИЧЕСКИХ СОЕДИНЕНИЯХ И БИОЛОГИЧЕСКИХ ОБЪЕКТАХ

Распознавание в текстах научных публикаций наборов символов, обозначающих названия химические соединения, обычно относят к классу задач распознавания так называемых поименованных сущностей. Мы будем использовать общепринятые аббревиатуры: NER – Named Entity Recognition, NE – Named Entity, CNER – Chemical Named Entity Recognition (распознавание названий химических соединений) [1].

Рассмотрим основные источники данных для извлечения информации о названиях и свойствах химических соединений, являющихся, главным образом, коллекциями текстов, которые принято называть корпусами. Они свободно доступны для загрузки и анализа.

На основе больших БД библиографической информации [6, 9] исследователями создаются корпусы, которые содержат размеченные под определённые задачи тексты различных тематик. Тексты таких корпусов могут содержать метки соответствия каждого слова определённому термину, например, наименованию ХС, белку, гену, какому-либо биологическому эффекту, а также терминам, обозначающим взаимодействия между объектами в тексте. Так, корпус CHEMDNER разработан консорциумом научных групп из тридцати четырёх различных организаций и представляет собой библиотеку текстов из более 10 тыс. рефератов БД NCBI PubMed, содержащих (на 10.08.2020) 84 355 наименований химических соединений, которые вручную аннотированы экспертами. Отбор текстов производили таким образом, чтобы они отражали специфику основных направлений химии [7].

Для поиска ассоциаций между ХС и индуцируемыми ими заболеваниями был создан корпус CDR (chemical-disease relation extraction), содержащий свыше 1,5 тыс. аннотированных экспертами публикаций с данными о конкретных заболеваниях, названиях химических соединений и ассоциациях между ними [10, 11].

Авторами BEL (biological expression language) разработан корпус для поиска ассоциаций между ХС и белками, заболеваниями и биологическими процессами (звеньями патогенеза конкретных заболеваний). На 10.08.2020 этот корпус содержит 11 тыс. кратких описаний ассоциаций «белок–ХС» и «заболевание–ХС» из более чем 6 тыс. текстов [12].

Корпус DrugNer [13] содержит тексты с размеченными названиями торговых наименований лекарственных препаратов, а также аннотации различных видов биологических эффектов, которые могут вызывать эти препараты. Тексты корпуса представлены в формате XML, один текст – один файл XML, разделены на предложения и пронумерованы. Для каждого предложения выделен фрагмент (фраза), который размечен в соответствии с содержанием.

Корпусы с данными по взаимодействию лекарственных препаратов (drug-drug interactions – DDI) [14] (версии 2013 г.) помимо торговых наименований препаратов, содержат данные о межлекарственном взаимодействии с белками человека, в том числе с ферментами, которые участвуют в биотрансформации химических соединений.

Все приведенные здесь корпусы доступны для загрузки и обработки. Помимо этих корпусов создаются новые – для решения конкретных научных задач, стоящих перед исследователями, поэтому общее количество корпусов постоянно растёт. В табл. 1 представлены основные корпусы, разработанные в последнее десятилетие для решения задач извлечения из научных публикаций наименований химических соединений и поиска ассоциаций между ними, а также между биологическими объектами и процессами.

Способы хранения массивов текстов отобранных для конкретных задач различны. Но на практике, как правило, используются реляционные базы данных. Преимущество такого способа – возможность доступа к хранилищу сразу нескольким пользователям, вариативность форматов используемых таблиц в рамках различных СУБД, что влияет на скорость обработки запросов.

Методы распознавания названий химических соединений CNER включают: (1) методы на основе словарей и/или систем грамматических и лексических правил [1, 8, 18, 19] и (2) методы машинного обучения, разработанные или модифицированные для задач автоматизированного анализа текстов.

Для реализации многих методов, использующих системы правил, применяют упорядоченные множества терминов (словари), для которых известно, что они являются Named Entity. Сочетание поиска термина в словаре с правилами распознавания определённой последовательности слов и терминов (паттерна) позволяет выявить NE. Основное ограничение методов, основанных на системе правил – их сравнительно узкая область применимости, поскольку невозможно обнаружить в тексте наименования, которые отсутствуют в словаре (например, редко используемые синонимы или аббревиатуры, введенные авторами статьи) или не встречаются в определённой последовательности слов, для распознавания которой созданы правила.

Таблица 1

Источники данных для поиска ассоциаций между химическими соединениями, белками и патологическими процессами у человека

Название корпуса	Аннотированные типы биологических объектов и их взаимодействия	Количество документов	Дата последнего обновления	Ссылка
CEMP	Наименования ХС, генов, белков в текстах патентов	Свыше 1 000 наименований химических субстанций	2017	[15]
BEL	ХС, заболевания, патологические процессы	Более 11 000 коротких описаний из более 6 000 текстов	2016	[12]
BioCreative V Chemical Disease Relation (BC5 CDR) corpus	Болезни, белки/гены, ХС	1 500	2015	[10, 11]
CHEMDNER	ХС, белки	Свыше 10 000	2013	[7]
DDIExtraction	Тексты, содержащие названия лекарственных препаратов	Свыше 1 000 текстов	2013	[14]
CRAFT	Тексты, содержащие разметку наименований ХС	67 текстов	2012	[16]
ИЮПАК training corpus	ХС	~1 500 абстрактов	2008	[17]
DrugNer	Лекарственные соединения	885 абстрактов	2008	[16]

Класс задач об извлечении данных из текстов относится к интеллектуальному анализу (text mining), в котором широко используют методы искусственного интеллекта и методы машинного обучения, базирующиеся на представлении исследуемого текста в виде специфических признаков, отражающих символы, слова или фразы.

ПРЕДСТАВЛЕНИЕ СЛОВ ПРИ АНАЛИЗЕ ТЕКСТА

В методах классификации, базирующихся на машинном обучении, используется набор признаков, характеризующих текст. Для распознавания категории, содержания и контекста фраз обычно применяется набор свойств отдельных слов, включающий, например, метку принадлежности к определённой части речи, к численным значениям (а также последовательность буквенных символов и цифр), заглавным (прописным) буквам (ортографические признаки текста) или последовательности символов конкретных слов (так называемые *n*-граммы), а также содержащий особенности морфологии конкретных слов (например, наличие или отсутствие префиксов, выделение минимальной смысловой части слова и т.п. – морфологические признаки). В некоторых методах применяются метки наличия или отсутствия в исследуемом тексте фиксированных смысловых сочетаний слов, характерных для языка (табл. 2), а также их комбинации [1].

Один из распространённых видов представления слов в задачах NER – признаки вида BIO, где B (*beginning*) – метка начала поименованной сущности (NE), I (*inside*) – метка принадлежности слова к NE, O (*outside*) – метка, означающая, что соответствующее слово не относится к NE. Помимо указанных признаков (см. табл. 2) применяются разметки слов, в которых содержится информация о том, относится ли данный термин к названию ХС в тексте, а также метки принадлежности одного слова или нескольких слов (сочетаний слов) к названию химического соединения [1].

В ряде исследований [1, 8, 19] показано, что применение специфических типов признаков, разработанных для конкретной задачи, является предпочтительным по сравнению с использованием сочетания признаков, которыми обычно представляют текст любой тематики.

Ещё один вариант признаков, применяемых при анализе текстов, – это векторное представление слов (word embeddings), в котором используется приведение каждого слова и отдельных фраз, последовательно генерируемых из предложения, в бинарный вектор-строку. Способы генерации векторов могут быть разнообразными, например, они могут отражать частоту слов в исследуемых корпусах. При этом подходе контекст, в котором находится конкретное слово, может быть учтён посредством перевода в векторы нескольких слов, расположенных в непосредственной близости с этим словом, и определения общего вектора, вычисленного посредством применения какой-либо функции (в простейшем случае - усреднением). Задача сравнения контекстов фраз в этом случае сводится к сравнению двух векторов, отражающих контекстное представление фраз.

Помимо признаков, отражающих морфологию слова или принадлежность его к части речи, иногда могут использоваться значения, являющиеся результатами классификации после применения иных инструментов, т. е. не определённые экспертом, а расчётные значения [1].

Таблица 2

Признаки, применяемые для извлечения названий химических соединений

Признаки	Описание	Пример	Ссылки
Морфологические	Отражают строение слова (суффиксы, префиксы и т.п.)	2-(Acetyloxy)benzoic acid = Acetyl + oxy + benz + oic + acid	[20-23]
Лемма	Нормализованная форма слова (в именительном падеже единственного числа)	Drugs -> drug	[24-25]
POS	Принадлежность к частям речи	2-(Acetyloxy)benzoic acid = noun	[20-25]
Ортографические	Первая буква слова соответствует её виду в предложении (заглавная или прописная), признаки отображают количество символов, относящихся к буквам, цифрам и знакам препинания в рассматриваемом термине	2-(Acetyloxy)benzoic acid = 2 (первый символ слова)+ 20 (количество букв) + 1 (количество цифр)	[20-23]
«форма слова»	Буквы в слове представлены: заглавная буква - A, строчная буква – a, цифры – O, все остальные символы - o	2-(Acetyloxy)benzoic acid = OooAaaaaaaaaoaaaaaaaoaaaa	[20-23]
BIO	Метка принадлежности слова B - началу наименования ХС, I - термину в составе наименования ХС, O - термину, не относящемуся к ХС вообще	2-(Acetyloxy)benzoic acid is aspirin = BI O B	[20-23]
NO/NE/S-NE/M-NE/E	Метки, соответствующие: NO – слову, не являющемуся названием ХС NE – начало наименования ХС; S/M – слово/множество слов являющееся названием ХС E – окончание термина, соответствующего названию ХС	2-(Acetyloxy)benzoic acid is aspirin = NE + ME + E-NE + NO + SE	[20-23]

МЕТОДЫ МАШИННОГО ОБУЧЕНИЯ В АНАЛИЗЕ ТЕКСТОВ

На использовании значительного количества примеров, для которых есть данные об их принадлежности к конкретной категории основаны методы машинного обучения. Примеры представляются в виде совокупности признаков (описание и примеры признаков приведены выше). Таким образом, по сравнению с методами, основанными на системе правил, методы машинного обучения обеспечивают более высокую степень абстракции и обладают более широкой областью применимости. К основным применяемым при анализе текстов методам машинного обучения можно отнести методы, основанные на теореме Байеса, методы опорных векторов, методы, основанные на построении искусственных нейронных сетей (ИНС). Одним из наиболее часто применяемых в анализе текстов и поиске взаимосвязей между терминами, извлечёнными в результате анализа, является метод Conditional Random Fields (CRF) [26]. Методы, которые являются его вариациями, основаны на оценке контекста, т. е., в формальном приближении, последовательности слов. Метод CRF основан на представлении текста в виде ненаправленного графа, где последовательности слов являются вершинами, соединёнными связями. Целевой переменной может быть принадлежность фрагмента текста к конкретной категории, либо принадлежность конкретного слова или словосочетания к определённому типу терминов (например, названию ХС). Для каждого полного подграфа этого целенаправленного графа определяется потенциальная функция, которая каждому возможному состоянию элементов подграфа ставит в соответствие вещественное число. Кроме признаков каждого слова при оценке целевой переменной во внимание принимается значение целевой переменной на предшествующем этапе оценки текущего значения потенциальной функции. Таким образом в модели может быть учтён контекст.

Методы, основанные на применении искусственных нейронных сетей для NER, могут базироваться на стандартной архитектуре или различных модифицированных вариантах искусственных нейронных се-

тей. Каждый искусственный нейрон в ИНС представлен нелинейной функцией от комбинации входных сигналов. Соответственно, каждый нейрон может иметь несколько «входов» и один «выход». В ИНС стандартной архитектуры содержится один слой для входных сигналов (параметров, input layer), один или несколько так называемых «скрытых» слоев и один слой, в котором формируются результаты рассчитанных значений целевой переменной. Подробно теоретические основы и применение ИНС (для задач «структура-свойство» ХС) рассмотрены в обзорах [27, 28]. В NER используются модифицированные варианты НС [29], например, свёрточные ИНС (convolutional neural networks, CNN), рекуррентные ИНС [30], а также сети «долгой-краткосрочной памяти» (Long-Short-Term Memory, LSTM) [31]. Идея метода основана на том, что процесс мышления тесно связан с процессами памяти. Поэтому ИНС необходимо хранить информацию, полученную в процессе обработки входного сигнала, и при некоторых условиях трансформировать эту информацию. Особенностью архитектуры этой ИНС является наличие в ней переменной, сохраняющей состояния сети, рассчитанные на предшествующем слое. Такой подход позволяет учитывать контекст при NER (например, учитывать род прилагательного, находящегося в непосредственной близости с NE).

Методы NER на основе CRF могут применяться также в комбинации с другими способами сравнения текстовых выражений, такими, как например, метод нечёткого поиска или метод на основе ИНС [32].

Результаты классификации оцениваются стандартными метриками: полнотой (*sensitivity*, *recall*), точностью – доля истинно-положительных относительно всех признанных положительными результатов (*precision*, *positive predictive value*) и величиной *F-score* (F_1-*score*) – средним гармоническим *precision* и *recall*.

В проблеме анализа текстов для извлечения данных о конкретных свойствах химических соединений, например, ассоциациях между ХС, белками или видами биологической активности, можно выделить две самостоятельные задачи: (1) извлечение данных о наименованиях ХС, белков, генов или биологических процессов и (2) автоматическое установление ассоциаций между найденными терминами. Далее последовательно рассмотрим примеры алгоритмов для решения каждой из этих задач.

ИЗВЛЕЧЕНИЕ ДАННЫХ О ХИМИЧЕСКИХ СОЕДИНЕНИЯХ ПРИ АНАЛИЗЕ ТЕКСТОВ

Методы, основанные на искусственных нейронных сетях различной архитектуры, CRF, а также на комбинации методов машинного обучения описаны в работах [22, 32–36].

Авторы [33] применяли для анализа текстов метод CRF с последующим пост-процессингом. Ими разработан список правил, который позволяет осуществлять отбор наиболее вероятных истинно-положительных результатов. Множество правил основано на совпадении числа открывающихся и закрывающихся скобок в извлечённой последовательности слов, которая отнесена алгоритмом к названию химического соединения. Точность (*precision*) распознавания названий ХС составила 83,71%. Следует отметить, что, хотя пост-процессинг выбранных алгоритмом последовательностей слов позволяет снизить число ложноположительных результатов и таким образом увеличить точность NER, тем не менее, он не позволяет снизить число ложноотрицательных результатов. Фильтрация терминов посредством применения системы правил может быть осуществлена как после получения результатов работы ИНС, так и на первом этапе (при подготовке данных) исследования [1, 23, 32].

Авторы работы [32] применили метод CRF в комбинации с алгоритмом CNN-BiLSTM. Отличительная особенность их подхода в том, что работа свёрточных ИНС основана на признаках последовательности пяти символов текста, в то время как на основе BiLSTM анализируется большее число символов; соответственно, может быть проанализирован контекст целых фраз и предложений. Окончательная оценка принадлежности нескольких слов к Named Entity производится с применением метода CRF, на вход которого подаются значения, полученные с использованием CNN-BiLSTM. При этом в цитируемом подходе распознавались не только NE химических соединений, но и названия белков, организмов, клеточных линий и тканей. В этой же работе отмечаются, что распознавание одновременно нескольких типов NE (например, одновременно ХС, белок, организм, тканевая принадлежность, клеточная линия) существенно не влияет на оценки точности для тестовой выборки в сравнении с распознаванием какого-либо одного типа NE. Так, значение *precision* 0,775; *recall* 0,587; F_1-*score* 0,668 было получено при CNER в случае, если модель была обучена для распознавания одновременно нескольких NE, и соответствующие значения точности составили *precision* 0,661; *recall* 0,681; F_1-*score* 0,671 при CNER без распознавания других терминов.

В работе [34] реализована многоуровневая искусственная нейронная сеть, в которой на вход подаются векторные представления слов и отдельных символов. Авторы использовали три типа архитектур ИНС: свёрточную, рекуррентную и распределённую. Результатом классификации стали данные о принадлежности последовательности символов (слова) конкретному классу из наименований ХС (например, тривиальное название (TRIVIAL), систематическая номенклатура (SYSTEMATIC) и т.п.). Сочетание трёх типов ИНС позволило извлечь термины, относящиеся к названию химического соединения, и определить их класс. Точность классификации на тестовой выборке составила: *precision* 0,886, *recal* 0,888, F_1-*score* 0,887.

Широкое развитие методов машинного обучения позволяет извлекать из текста названия химических соединений, биологических объектов, процессов, а также определять ассоциации между ними. В то же время, поскольку при работе алгоритмов машинного обучения возможны ошибки ложно распознаваемых объектов [1], исследователи вынуждены вносить ограничения на работу алгоритма посредством построения системы правил, ранжирующих результаты

классификации, полученные методами машинного обучения. Таким образом, возникают комбинированные (или гибридные) подходы, которые основаны на методах машинного обучения вместе с применением совокупности правил или с осуществлением поиска в словарях терминов ХС.

В работе [35] авторы использовали модифицированный алгоритм BiLSTM-CRF, в котором совокупность признаков для отдельных слов подвергалась нескольким последовательным преобразованиям на каждом из этапов которых последовательно применяли функцию гиперболического тангенса, алгоритм CRF для каждого слова из предложения, причём на вход алгоритма CRF подавали последовательность из нескольких слов, которая была получена (1) при движении от начала к одному выбранному слову в предложении и (2) от конца к этому же слову, рассчитанную с учётом слов в предложении, предшествующих исследуемому. На конечном этапе каждое слово подавали на вход в алгоритм CRF, и затем оценивали результаты классификации. Максимальные параметры точности CNER для корпуса CHEMNDER при этом подходе составили: *precision* 0,917, *recall* 0,904.

Аналогично комбинированный подход реализован в методе LSTMVoter [36]. В LSTMVoter использован метод CRF в сочетании с искусственной нейронной сетью, основанной на архитектуре LSTM. На вход, помимо стандартных параметров, также подавались оценки классификации, рассчитанные другими методами CNER, разработанными ранее (табл. 3).

Среди комбинированных систем CNER также можно отметить ChemSpot [22]. В этом алгоритме метод CNER на основе CRF сочетали с поиском терминов, найденных в словаре Международного союза теоретической и прикладной химии IUPAC (International Union of Pure and Applied Chemistry). Авторы ChemSpot подчёркивают, что такой способ дает возможность однозначно определять соответствие найденному термину в словаре ИЮПАК, что, в свою очередь, позволяет избегать ошибок, опечаток, неточного названия в тексте при распознавании и сохранении термина ИЮПАК. Такой подход сочетает в себе элементы машинного обучения и метода, основанного на системе правил.

Подходы, используемые для поиска названий химических соединений, во многих случаях основаны на применении методов машинного обучения и искусственных нейронных сетей для интеграции результатов классификации, в том числе, полученных другими методами. Учитывая, что задача CNER в текстах научных публикаций является достаточно трудно формализуемой, машинное обучение может применяться совместно с подходами, основанными на системе словарей или системе правил фильтрации терминов, не имеющих отношения к названиям химических соединений (пост-процессинг).

Таблица 3

Примеры методов, реализующих поиск названий химических соединений в текстах научных публикаций

Цель метода	Корпус	Представление слов	Метод обучения	Средняя точность распознавания	Ссылка
CNER	CHEMNDER	Группировка символов на последовательность букв и цифр	Искусственные нейронные сети	Около 0,89	[34]
CNER	CHEMNDER CEMP	Разметка BIO	Комбинированный метод (LSTM-ANN-CRF)	0,89	[36]
NER ХС в совокупности с другими NE (белки, организмы и т.п.)	BIOCREATIVE (BIO-ID)	Разметка BIO, токенизация, стандартизация синонимов	Комбинированный метод (CNN-LSTM-CRF)	0,67	[32]
CNER	CHEMNDER	Комбинация признаков: - векторные представления - *n*-граммы - метки частей речи (POS tags)	Комбинированный метод (CRF)	0,90	[35]
CNER	MEDLINE	Последовательность из нескольких символов (n-граммы), коллокации (совместная встречаемость терминов), POS-tags	Conditional random fields	0,87	[33]
CNER	MEDLINE	Разметка слов по принадлежности к терминам ХС	Conditional random fields + поиск в словаре ИЮПАК	0,89	[22]

ПОИСК АССОЦИАЦИЙ МЕЖДУ ИЗВЛЕЧЁННЫМИ НАЗВАНИЯМИ ХИМИЧЕСКИХ СОЕДИНЕНИЙ, БЕЛКАМИ И БИОЛОГИЧЕСКИМИ ПРОЦЕССАМИ

Автоматическое распознавание CNER может быть первым этапом в экстракции данных о взаимосвязи химических соединений с биологическими эффектами, которые они могут потенциально проявлять при условии, что автоматический поиск проводят в научных публикациях, в которых содержится экспериментальное подтверждение взаимодействия ХС с белками-мишенями или наличия определённого вида биологической активности. Разработка автоматических методов экстракции данных о ХС и их биологических эффектах или других свойствах может способствовать созданию обучающих выборок для методов анализа взаимосвязи «структура – активность» (Structure – Activity Relationship, SAR). Данные об экспериментально подтверждённых видах биологической активности химических соединений могут быть применены для поиска дополнительных фармакотерапевтических свойств лекарственных препаратов (Drug Repurposing).

Выявление различных биологических эффектов и/или взаимодействий между ХС и белками в литературе реализовано несколькими методами: (1) поиск названия активности/имени белка; (2) система правил, где указаны основные термины и характеристика взаимодействий белок-лиганд, описываемая с помощью этих терминов; (3) поиск на основе правил определённых «паттернов» – закодированных в определённой последовательности обобщённых терминов; (4) классификация фраз из текста на относящиеся или не относящиеся к описанию взаимодействий ХС с белками на основе методов машинного обучения; (5) определение совместной частоты встречаемости двух или нескольких терминов (например, названия лекарственного препарата и определённой биологической активности) в тексте. Далее приведем примеры установления ассоциаций между химическими соединениями и их биологическими свойствами.

В исследовании [37] предложения классифицировали на две категории: (1) содержащие описание взаимодействий между ХС и белком и (2) не содержащие такого описания. Обучающая выборка состояла из 1 632 аннотаций статей из БД научных публикаций NCBI PubMed. Предложения статьи обучающей выборки были проаннотированы и выбраны фразы, в которых содержались упоминания белков и химических соединений (пары NE), и фразы, в которых, дополнительно к парам NE, содержались упоминания о типе взаимодействий между белком и ХС. В качестве признаков использовали принадлежность аннотированных фраз к одному из семи типов взаимного расположения слов, описывающих термины (ХС, белок) в тексте, расстояние между терминами, выраженное в количестве слов, количество слов в предложении, наличие или отсутствие между терминами других слов биомедицинской тематики, признанных авторами подхода «значимыми» для выявления пар и триплетов NE, описывающих взаимодействие между ХС и белками. Классификация предложений основана на применении нескольких методов машинного обучения, включая деревья решений, наивный Байесов подход, логистическую регрессию, линейный дискриминантный анализ. Для определения принадлежности предложения к категории (1) или (2) было выявлено пересечение результатов отдельных классификаторов. Авторы [37] отмечают, что значения точности, полученные при реализации указанного метода (Precision 0,63; Recall: 0,51), сопоставимы с результатами, основанными на глубоком обучении искусственных нейронных сетей.

К настоящему времени разработаны и реализованы в виде веб-сервисов алгоритмы по выявлению потенциальных биологических эффектов для низкомолекулярных химических соединений и лекарственных препаратов. Большинство из этих алгоритмов нацелено на выявление одной или нескольких типов биологической активности, которые могут быть взаимосвязаны, например, токсичность и воздействие на ферменты метаболизма. Алгоритм LimTox (Literature Mining for Toxicology), реализованный в виде веб-сервиса), направлен на поиск биологических эффектов лекарственных препаратов, ассоциированных с гепатотоксичностью и взаимодействием с ферментами семейства цитохромов Р450. Поиск можно проводить и для других видов токсичности, таких как нефротоксичность, гепатотоксичность, кардиотоксичность [18].

В алгоритме, реализованном в ChemoText [38], существует возможность поиска терминов MeSH (Medical Subject Headings – это тезаурус словаря, контролируемый NLM (National Library of Medicine), используемый для индексации статей для PubMed) [39], ассоциированных с тремя типами NE: (1) низкомолекулярное химическое соединение; (2) белок; (3) заболевание. В ChemoText существует возможность поиска по терминам MeSH, анализа совместной встречаемости терминов, задаваемых пользователем, в списке терминов MeSH, а также построения с применением терминов MeSH семантических карт. Результаты поиска могут быть применены для (1) отбора публикаций, в которых исследуются конкретные низкомолекулярные ХС либо интересующий исследователя белок-мишень, и (2) для оценки ассоциаций между конкретными белками и заболеваниями и поиска препаратов, потенциально применимых для терапии этих заболеваний. Такой «двунаправленный» поиск и возможность отбирать публикации, в которых рассматриваются возможные звенья патогенеза заболеваний, а также химические соединения, предположительно ассоциированные с терапией заболевания, является преимуществом метода. В качестве ограничений можно отметить, что поиск производится только по MeSH терминам, исключая информацию из полного текста статьи. Кроме того, что при таком подходе теряется часть информации, известно, что термины MeSH доступны не для всех публикаций (они отсутствуют примерно в 10 % всех публикаций) в БД PubMed, что может приводить к ограничению множества анализируемых статей.

Подходы, объединяющие данные о химических соединениях и их возможных биологических эффектах, включающих взаимодействие с белками-мишенями, могут быть основаны и на сопоставлении названий белков-мишеней, найденных в тек-

стах публикаций, с таковыми из БД, содержащих информацию о них.

В работе Е. А. Пономаренко и соавторов были построены семантические сети для множества белков, ассоциированных с метаболическими путями [40]. Для построения семантических сетей авторы применяли выборку публикаций, которые были найдены в PubMed при добавлении названия белка в строку запроса (запросы сгенерированы автоматически). При выполнении поискового запроса были выявлены релевантные публикации для каждого белка. Затем определено пересечение множеств релевантных публикаций для пар исследуемых белков. На основании количества публикаций, релевантных для каждой пары исследуемых белков одновременно, и суммарного количества публикаций, релевантных хотя бы одному белку, были рассчитаны коэффициенты подобия (коэффициент Танимото). На основании функции, зависящей от значений коэффициента Танимото, были сформированы группы белков. Впоследствии были показаны общие функции для белков, входящих в состав одной группы, согласно предложенному методу. Разработанный алгоритм может быть применён и для выявления схожих свойств химических соединений.

ЗАКЛЮЧЕНИЕ

Автоматическое распознавание названий химических соединений в текстах научных публикаций позволяет решать ряд задач: автоматизированное пополнение баз данных о схемах химического синтеза, синтетической доступности ХС, в том числе, с требуемым видом биологической активности; поиск новых видов фармакологической активности для зарегистрированных лекарственных препаратов, а также возможных побочных эффектов, оценку вероятных межлекарственных взаимодействий.

Вместе с тем, CNER является достаточно трудоёмкой, сложно формализуемой задачей, поскольку тексты на любом языке имеют вариабельную структуру. Поэтому применяется ряд признаков, описывающих как непосредственно отдельные фрагменты текста (слова, символы), так и разметку, позволяющую учесть контекст. Помимо специфических признаков, контекст обычно учитывается при построении модели (например, в методе CRF возможно учитывать контекст; существуют варианты искусственных нейронных сетей, в которых предшествующее состояние сохраняется при прохождении по этим сетям). Точность распознавания (*precision*) названий химических соединений для большинства разрабатываемых методов извлечения наименований ХС находится в диапазоне 0,67 – 0,90. Точность CNER и других терминов биомедицинской тематики может быть существенно повышена, если в научных публикациях указывать принадлежность биологического или химического термина к классификации Bioassay Ontology [41] или Chemical Information Ontology [42]. Введение ограничений на обязательный конкретный формат названия химического соединения в статьях, например, формат ИЮПАК, позволяет увеличивать точность распознавания для задач NE. Помимо этого, разработка новых признаков для представления слов и применение дополнительных методов отбора релевантных публикаций для класса задач CNER может способствовать улучшению точности распознавания ассоциаций между названиями ХС и их вероятной биологической активностью, и, как следствие, увеличению объёма и повышению качества данных о свойствах химических соединений

СПИСОК ЛИТЕРАТУРЫ

1. Krallinger M., Rabal O., Lourenço A., Oyarzabal J., Valencia A. Information Retrieval and Text Mining Technologies for Chemistry // Chemical Reviews. – 2017. – Vol. 117, № 12. – P. 7673–7761.
2. Przybyła P., Shardlow M., Aubin S., Bossy R., Eckart de Castilho R., Piperidis S., McNaught J., Ananiadou S. Text mining resources for the life sciences // Database. – 2016. – Vol. 2016 (baw145), P. 1-30.
3. Oellrich A., Gkoutos G.V., Hoehndorf R., Rebholz-Schuhmann D. Quantitative comparison of mapping methods between Human and Mammalian Phenotype Ontology // Journal of Biomedical Semantics. – 2012. – Vol. 3, № S2/S1. – P. 1-10.
4. O'Mara-Eves A., Thomas J., McNaught J., Miwa M., Ananiadou S. Using text mining for study identification in systematic reviews: a systematic review of current approaches // Systematic Reviews. – 2015. – Vol. 4, № 5. – P. 1-22.
5. Smink W.A.C., Fox J.-P., Tjong Kim Sang E., Sools A.M., Westerhof G.J., Veldkamp B.P. Understanding Therapeutic Change Process Research Through Multilevel Modeling and Text Mining // Frontiers in Psychology. – 2019. – Vol. 10. – P. 1186.
6. PubMed. – URL: https://pubmed.ncbi.nlm.nih.gov/
7. Krallinger M., Rabal O., Leitner F., Vazquez M., Salgado D., Lu Zh., Leaman R., Lu Y., Ji D., Lowe D.M., Sayle R. A., Batista-Navarro R.Th., Rak R., Huber T., Rocktäschel T., Matos S., Campos D., Tang B., Xu H., Munkhdalai T., Ryu K.H., Ramanan S.V., Nathan S., Žitnik S., Bajec M., Weber L., Irmer M., Akhondi S.A., Kors J.A., Xu Sh., An X., Sikdar K.U., Ekbal A., Yoshioka M., Dieb Th.M., Choi M., Verspoor K., Khabsa M., Giles C.L., Liu, H., Komandur Ravikumar K.E., Lamurias A., Couto F.M., Dai H.-D., Tzong-Han Tsai R., Ata C., Can T., Usié A., Alves R., Segura-Bedmar I., Martínez P., Oyarzabal J., Valencia A. The CHEMDNER corpus of chemicals and drugs and its annotation principles // Journal of Cheminformatics. – 2015. – Vol. 7, № S2. – P. 2-17.
8. Akhondi S.A., Hettne K.M., van der Horst E., van Mulligen E.M., Kors J.A. Recognition of chemical entities: combining dictionary-based and grammar-based approaches // Journal of Cheminformatics. – 2015. – Vol. 7 (Suppl 1: S6) – P. 1-10.
9. NCBI. – URL: https://www.ncbi.nlm.nih.gov/mesh/

10. Li J., Sun Y., Johnson R.J., Sciaky D., Wei C.-H., Leaman R., Davis A.P., Mattingly C.J., Wiegers T.C., Lu Z. BioCreative V CDR task corpus: a resource for chemical disease relation extraction // Database. – 2016. – Vol. 2016 (baw086). – P. 1-10.
11. Wei C.-H., Peng Y., Leaman R., Davis A.P., Mattingly C.J., Li J., Wiegers T.C., Lu Z. Assessing the state of the art in biomedical relation extraction: overview of the BioCreative V chemical-disease relation (CDR) task // Database. – 2016. – Vol. 2016 (baw032). P. 1-8.
12. Madan S., Szostak J., Komandur Elayavilli R., Tsai R.T.-H., Ali M., Qian L., Rastegar-Mojarad M., Hoeng J., Fluck J. The extraction of complex relationships and their conversion to biological expression language (BEL) overview of the BioCreative VI (2019) BEL track // Database. – 2019. – Vol. 2019 (baz084). – P. 1-17.
13. Martínez V., Navarro C., Cano C., Fajardo W., Blanco A. DrugNet: Network-based drug–disease prioritization by integrating heterogeneous data // Artificial Intelligence in Medicine. – 2015. – Vol. 63, № 1. – P. 41-49.
14. Herrero-Zazo M., Segura-Bedmar I., Martínez P., Declerck T. The DDI corpus: An annotated corpus with pharmacological substances and drug–drug interactions // Journal of Biomedical Informatics. – 2013. – Vol. 46, № 5. – P. 914-920.
15. Pérez-Pérez M., Rabal O., Pérez-Rodríguez G., Vazquez M., Fdez-Riverola F., Oyarzabal J., Valencia A., Lourenço A., Krallinger M. Evaluation of chemical and gene/protein entity recognition systems at BioCreative V.5: the CEMP and GPRO patents tracks // Proceedings of the BioCreative. Vers. 5. Challenge Evaluation Workshop. – 2017. – P. 11-18. – URL: https://biocreative.bioinformatics.udel.edu/media/store/files/2017/ BioCreative_V5_paper2.pdf
16. Bada M., Eckert M., Evans D., Garcia K., Shipley K., Sitnikov D., Baumgartner Jr.W.A., Cohen B., Verspoor K., Blake J.A., Hunter L.E. Concept annotation in the CRAFT corpus // BMC Bioinformatics. – 2012. – Vol. 13, № 161. – P. 1-10.
17. Kola´rik C., Klinger R., Friedrich C.M., Hofmann-Apitius M., Fluck J. Chemical Names: Terminological Resources and Corpora Annotation // Workshop on Building and Evaluating Resources for Biomedical Text Mining (6th edition of the Language Resources and Evaluation Conference). – Marrakech (Morocco), 2008. – P. 51-58. – URL: https://pub.uni-bielefeld.de/record/2603498
18. Cañada A., Capella-Gutierrez S., Rabal O., Oyarzabal J., Valencia A., Krallinger M. LimTox: a web tool for applied text mining of adverse event and toxicity associations of compounds, drugs and genes // Nucleic Acids Research. – 2017. – Vol. 45, № W1. – P. W484-W489.
19. Swain M.C., Cole J.M. ChemDataExtractor: A Toolkit for Automated Extraction of Chemical Information from the Scientific Literature // Journal of Chemical Information and Modeling. – 2016. – Vol. 56, № 10.– P. 1894-1904.
20. Batista-Navarro R., Rak R., Ananiadou S. Optimising chemical named entity recognition with pre-processing analytics, knowledge-rich features and heuristics // Journal of Cheminformatics. – 2015. – Vol. 7 (Suppl 1: S6). – P. 1-13.
21. Leaman R., Khare R., Lu Z. Challenges in clinical natural language processing for automated disorder normalization // Journal of Biomedical Informatics. – 2015. – Vol. 57.–- P. 28-37.
22. Rocktäschel T., Weidlich M., Leser U. ChemSpot: a hybrid system for chemical named entity recognition // Bioinformatics. – 2012. – Vol. 28, № 12. – P. 1633-1640.
23. Campos D., Bui Q.-C., Matos S., Oliveira J.L. TrigNER: automatically optimized biomedical event trigger recognition on scientific documents // Source Code for Biology and Medicine. – 2014. – Vol. 9, №1. – P. 1.
24. Lu Z., Hirschman L. Biocuration workflows and text mining: overview of the BioCreative 2012 Workshop Track II // Database. – 2012. – Vol. 2012 (bas043). – P. 1-6.
25. Liu H., Christiansen T., Baumgartner W.A., Verspoor K. BioLemmatizer: a lemmatization tool for morphological processing of biomedical text // Journal of Biomedical Semantics. – 2012. – Vol. 3, №3. – P. 1-29.
26. Song H.-J., Jo B.-C., Park C.-Y., Kim J.-D., Kim Y.-S. Comparison of named entity recognition methodologies in biomedical documents // BioMedical Engineering OnLine. – 2018. – Vol. 17 (Suppl 2). – P. 158-192.
27. Halberstam N.M., Baskin I.I., Palyulin V.A., Zefirov N.S. Neural networks as a method for elucidating structure–property relationships for organic compounds // Russian Chemical Reviews. – 2003. – Vol. 72, № 7. – P. 629-649.
28. Baskin I.I., Madzhidov T.I., Antipin I.S., Varnek A.A. Artificial intelligence in synthetic chemistry: achievements and prospects // Russian Chemical Reviews. – 2017. – Vol. 86, №11. – P. 1127-1156.
29. Cho H., Lee H. Biomedical named entity recognition using deep neural networks with contextual information // BMC bioinformatics. – 2019. – Vol. 20, №1. – P. 735-746.
30. Maheswaranathan N., Williams A.H., Golub M.D., Ganguli S., Sussillo D. Reverse engineering recurrent networks for sentiment classification reveals line attractor dynamics // Advances in Neural Information Processing Systems. – 2019. – Vol. 32. – P. 15696-15705.
31. Li Z., Gurgel H., Dessay N., Hu L., Xu L., Gong P. Semi-Supervised Text Classification Framework: An Overview of Dengue Landscape Factors and Satellite Earth Observation // International Journal of Environmental Research and Public Health. – 2020. – Vol. 17, №12. – P. 4509-4538.
32. Kaewphan S., Hakala K., Miekka N., Salakoski T., Ginter F. Wide-scope biomedical named entity recognition and normalization with

CRFs, fuzzy matching and character level modeling // Database. – 2018. – Vol. 2018 (bay096). – P. 1-10
33. Campos D., Matos S., Oliveira J.L. A document processing pipeline for annotating chemical entities in scientific documents // Journal of Cheminformatics. – 2015. – Vol. 7 (Suppl 1: S7). – P.1-10.
34. Korvigo I., Holmatov M., Zaikovskii A., Skoblov M. Putting hands to rest: efficient deep CNN-RNN architecture for chemical named entity recognition with no hand-crafted rules // Journal of Cheminformatics. – 2018. – № 1. – P. 28.
35. Luo L., Yang Z., Yang P., Zhang Y., Wang L., Lin H., Wang J. An attention-based BiLSTM-CRF approach to document-level chemical named entity recognition // Bioinformatics. – 2018. – Vol. 34, № 8. – P. 1381-1388.
36. Hemati W., Mehler A. LSTMVoter: chemical named entity recognition using a conglomerate of sequence labeling tools // Journal of Cheminformatics. – 2019. – Vol. 11, № 3. – P. 1-7.
37. Lung P.-Y., He Z., Zhao T., Yu D., Zhang J. Extracting chemical–protein interactions from literature using sentence structure analysis and feature engineering // Database.–- 2019. – Vol. 2019 (bay138). – P. 1-8.
38. Capuzzi S.J., Thornton T.E., Liu K., Baker N., Lam W.I., O'Banion C.P., Muratov E.N., Pozefsky D., Tropsha A. Chemotext: A Publicly Available Web Server for Mining Drug–Target–Disease Relationships in PubMed // Journal of Chemical Information and Modeling. – 2018. – Vol. 58, № 2. – P. 212-218.
39. Mao Y., Lu Z. MeSH Now: automatic MeSH indexing at PubMed scale via learning to rank // Journal of Biomedical Semantics. – 2017. – Vol. 8, № 1. – P. 15-24.
40. Пономаренко Е.А., Лисица А.В., Ильгисонис Е.В., Арчаков А.И. Создание семантических сетей белков с использованием PUBMED/MEDLINE // Молекулярная Биология. – 2010. – Т. 44, № 1. – С. 152-161.
41. Vempati U.D., Schürer S.C. Development and Applications of the Bioassay Ontology (BAO) to Describe and Categorize High-Throughput Assays // Assay Guidance Manual / eds. S. Markossian, G.S. Sittampalam, A. Grossman, et al. – Bethesda: Eli Lilly & Company and the National Center for Advancing Translational Sciences, 2004. – P.1045-1069.
42. Hastings J., Chepelev L., Willighagen E., Adams N., Steinbeck Ch., Dumontier M. The Chemical Information Ontology: Provenance and Disambiguation for Chemical Data on the Biological Semantic Web // PLoS ONE. – 2011. – Vol. 6, № 10. – P. e25513.

Материал поступил в редакцию 31.08.20.

Сведения об авторах

БИЗЮКОВА Надежда Юрьевна – лаборант Научно-исследовательского института биомедицинской химии имени В.Н. Ореховича, студентка шестого курса специальности «Медицинская кибернетика» Российского национального исследовательского медицинского университета имени Н.И. Пирогова, Москва
e-mail: nad.smol@gmail.com

ТАРАСОВА Ольга Александровна – кандидат биологических наук, научный сотрудник Научно-исследовательского института биомедицинской химии имени В.Н. Ореховича
e-mail: olga.a.tarasova@gmail.com

РУДИК Анастасия Владимировна – кандидат биологических наук, старший научный сотрудник Научно-исследовательского института биомедицинской химии имени В.Н. Ореховича
e-mail: rudik_anastassia@mail.ru

ФИЛИМОНОВ Дмитрий Алексеевич – кандидат физико-математических наук, ведущий научный сотрудник Научно-исследовательского института биомедицинской химии имени В.Н. Ореховича
e-mail: dmitry.filimonov@ibmc.msk.ru

ПОРОЙКОВ Владимир Васильевич – доктор биологических наук, кандидат физико-математических наук, член-корреспондент РАН, профессор, главный научный сотрудник, заведующий отделом биоинформатики Научно-исследовательского института биомедицинской химии имени В.Н. Ореховича
e-mail: vladimir.poroikov@ibmc.msk.ru;
vvp1951@yandex.ru

ВИНИТИ РАН

Центр научно-информационного обслуживания

Информационные услуги, предоставляемые ЦНИО ВИНИТИ 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 ВИНИТИ (разработка ВИНИТИ), 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 ВИНИТИ РАН www.viniti.ru

К услугам пользователей – **Электронный Каталог ВИНИТИ 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 ВИНИТИ РАН.

Подробную информацию Вы можете получить:
Адрес: 125190, Россия, г. Москва, ул. Усиевича, 20, ВИНИТИ РАН;
Телефоны: 499-155-42-17, 499-155-42-43;
E-mail: cnio@viniti.ru

ВНИМАНИЮ ЧИТАТЕЛЕЙ!

ИЗДАНИЕ УДК

УНИВЕРСАЛЬНАЯ ДЕСЯТИЧНАЯ КЛАССИФИКАЦИЯ

АЛФАВИТНО-ПРЕДМЕТНЫЙ УКАЗАТЕЛЬ

в 2-х томах

Алфавитно-предметный указатель (АПУ) к 4-му полному изданию УДК на русском языке:

Том I содержит АПУ от буквы А до Н;

Том II содержит АПУ от буквы М до Я и указатель латинских наименований к классам УДК 56 Палеонтология, 57 Биологические науки, 58 Ботаника, 49 Зоология, 61 Медицинские науки.

АПУ содержит около 100 000 понятий, представленных в полных таблицах УДК.

При его составлении были учтены изменения, опубликованные в Выпусках № 1 – 6 «Изменения и дополнения к УДК»

Для подписки необходимо направить заявку для оформления счета по адресу:

125190, Россия, Москва, ул. Усиевича, 20, ВИНИТИ РАН

Телефоны: 499 155-42-85, 499 151-78-61

E-mail: feo@viniti.ru

http://www.udcc.ru