ФЕДЕРАЛЬНОЕ АГЕНТСТВО НАУЧНЫХ ОРГАНИЗАЦИЙ

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

НАУЧНО • ТЕХНИЧЕСКАЯ ИНФОРМАЦИЯ

Серия 1. ОРГАНИЗАЦИЯ И МЕТОДИКА ИНФОРМАЦИОННОЙ РАБОТЫ

ЕЖЕМЕСЯЧНЫЙ НАУЧНО-ТЕХНИЧЕСКИЙ СБОРНИК

Издается с 1961 г. **№ 11** **Москва 2017**

ОБЩИЙ РАЗДЕЛ

УДК 005.94 : 004.9

М.Р. Биктимиров, Б.С. Есенькин, П.А. Зотов, Е.Б. Ногина, Я.Л. Шрайберг

Инфраструктура знаний – важнейший компонент цифровой экономики

Создание инфраструктуры знаний – идея, объединившая ведущие организации книжной и информационной сферы для формирования государственной системы, аккумулирующей накопленные человеком знания и обеспечивающей основные процессы управления знаниями: извлечение, хранение, идентификацию, систематизацию и классификацию информации, а также достоверность знаний и предоставление доступа к ним.

Ключевые слова: *управление знаниями, инфраструктура, цифровая экономика, индустрия знаний, информационный поток, классификация, идентификатор, оцифровка, электронная библиотека, издание, каталог, интернет, доступ, интеллектуальная собственность*

По определению ЮНЕСКО знаниями следует считать умения человека эффективно использовать накопленную информацию. При этом ценность самих знаний заключается в их достатке и полноте, т.е. известный тезис о прямой зависимости владения миром и владения информацией трансформировался в другой: миром владеет тот, кто умеет информацию эффективно использовать и владеет многими знаниями. Управление знаниями осуществляется посредством неких систематических процессов, благодаря кото-

рым создаются, извлекаются, собираются, сохраняются, распределяются и применяются знания как основные элементы интеллектуального капитала, необходимые для успешной деятельности человека. Это процессы:

- поиска и извлечения существующих знаний;
- структуризации, классификации и идентификации знаний;
- создания новых знаний;
- передачи и обеспечения доступа (в т.ч. авторизованного) к знаниям;
- поддержания целостности и достоверности знаний;
- использования и воплощения знаний в продуктах и услугах.

Комплексом взаимосвязанных систем и процессов, обеспечивающих функционирование и развитие средств взаимодействия субъектов в сфере знаний, как раз и является инфраструктура знаний. Таким образом, мы говорим сегодня об инфраструктуре знаний как о совокупности средств управления знаниями, объединяющей различные субъекты производства и распространения знаний из всего диапазона общественной жизни, в первую очередь, из области науки, культуры и просвещения.

Напомним некоторые положения из выступления Президента России на Петербургском международном экономическом форуме, где он говорил, что формирование цифровой экономики – это вопрос национальной безопасности и независимости России и выделил четыре главных направления: «…Первое – необходимо сформировать принципиально новую, гибкую нормативную базу для внедрения цифровых технологий во все сферы жизни. При этом все решения должны приниматься с учётом обеспечения информационной безопасности государства, бизнеса и граждан. Второе – государство окажет поддержку… носителям разработок и компетенций в сфере цифровых технологий, имеющих так называемый сквозной межотраслевой эффект. Это обработка и анализ больших массивов данных, искусственный интеллект и нейротехнологии, технологии виртуальной и дополненной реальности и ряд других. Третье – с участием государства и частного бизнеса будем создавать опорную инфраструктуру цифровой экономики, в том числе безопасные линии связи и центры обработки данных. …Это должна быть инфраструктура, основанная на самых передовых технологиях и разработках. Четвёртое – намерены кратно увеличить выпуск специалистов в сфере цифровой экономики, а, по сути, нам предстоит решить более широкую задачу, задачу национального уровня – добиться всеобщей цифровой грамотности…»

Как же связана инфраструктура знаний с цифровой экономикой в такой постановке проблемы? Кажется, ответ очевиден: цифровая экономика – это прежде всего экономика знаний. Ведь в современном мире конкурентным преимуществом всё больше становится не столько выгодная рыночная позиция, сколько владение широким набором необходимых, порой уникальных знаний, т.е. суть экономики знаний заключается в эффективном использовании извлеченных из больших информационных потоков знаний как экономических, политических или социальных активов. Извлеченные, а затем структурированные и сохраненные знания образуют компетенции, которые, в свою очередь, служат основой для производства товаров и предоставления услуг.

В настоящий момент получили развитие информационно-коммуникационные и когнитивные технологии, позволяющие организовать извлечение и хранение всех видов знаний, наряду с печатной формой, в электронном виде, т.е. весь контент может храниться в однородном электронном виде независимо от того, в какой форме он был первоначально издан. Новые технологии позволяют организовать:

- распределенное хранение знаний в электронном виде, при этом за хранение может отвечать неограниченное количество организаций и физических лиц;
- быструю и практически бесплатную транспортировку знаний на любые расстояния;
- быстрый автоматизированный поиск и доступ к знаниям, хранящимся в электронном виде. При этом для пользователя практически нет географических ограничений.

Цифровые технологии стирают ранее существовавшие четкие функциональные роли участников индустрии знаний. Практически каждый участник может играть одновременно роль автора, издателя, продавца, потребителя и хранилища знаний. Роль участника в каждый конкретный момент определяется характером прав, которыми участник обладает на конкретный объект интеллектуальной собственности, и способом, которым участник собирается использовать данный объект.

Широкое применение цифровых технологий кардинально упрощает неправомерное использование опубликованного знания и требует применения специальных процедур, направленных на обеспечение прав на интеллектуальную собственность при операциях с издательской продукцией в электронной форме.

Цифровые технологии формируют тенденцию к еще большему, в перспективе, снижению тиражей бумажных изданий и в целом – большей диверсификации и адресности номенклатуры издательской продукции по целевым группам потребителей. Это делает вопросы внедрения общепризнанных стандартов оформления публикаций, их производства и обращения особенно актуальными. Весь изданный за всю историю контент, включая издания до начала цифровой эпохи, может и должен существовать помимо бумажной и в электронной форме. Таким образом, все имеющиеся и опубликованные на настоящий момент знания так или иначе будут сосредоточены в цифровых хранилищах различных архивных и библиотечных ресурсов.

Стирание четких различий в функциональных ролях участников индустрии знаний, формирование новых способов потребления знаний и доступа к ним создает ситуацию, при которой практически любой участник может напрямую взаимодействовать с любым другим участником. При этом, роли партнеров такого взаимодействия могут быть различными в ка-

ждом конкретном случае. В таких обстоятельствах роль участника определяется не столько его местом и ролью в производственно-технологической цепочке (автор – издатель – распространитель – библиотекарь – потребитель), сколько его юридическим статусом в отношении прав собственности на то или иное произведенное и опубликованное знание. Такая свобода взаимодействия создаст условия для ранее не существовавшей интенсивности обращения знаний. Будет сформирована качественно новая информационная среда, позволяющая говорить о реальном переходе к стадии «общества знаний» и созданию нового культурного и технологического уклада.

Для того чтобы новые возможности позволили полностью раскрыть ранее недоступный потенциал и не превратили взаимоотношение участников в хаос, их взаимодействие должно быть упорядочено. Изменение вида основного продукта, а также изменение ролей участников индустрии знаний потребует принципиально новой «информационной инфраструктуры», или как мы её назвали, инфраструктуры знаний. Эта инфраструктура необходима для создания условий, при которых можно будет многократно увеличить объем реализуемых на рынке прав на интеллектуальную собственность.

В составе инфраструктуры знаний наряду со специализированными тематическими каталогами может быть сформирован единый универсальный каталог. Специализированные каталоги совместимы и связаны ссылками с универсальным каталогом. Необходимы также сертифицированные хранилища электронного содержания. Сеть уполномоченных хранилищ оригинальных образцов содержания всех форм изданий должна обеспечивать:

- достоверные ссылки на контент в Интернете – это будут страховые фонды знаний различной направленности;
- верификацию изданий, находящихся на рынке, на предмет их тождественности оригинальной версии.

Для этого потребуется специализированное программное обеспечение, гарантирующее контроль и сопровождение процесса обращения прав на опубликованную интеллектуальную собственность.

Конечно, в России уже существует достаточно большое количество ресурсов, аккумулирующих разнообразную информацию и извлеченные из нее знания. Благодаря системе обязательного экземпляра, существующей в России много лет, Российская книжная палата (РКП) комплектует литературой фонды крупнейших библиотек страны. РКП ведет банк данных государственной библиографии России, который может быть основой для создания единого государственного универсального библиографического каталога, способного обеспечивать в условиях новых технологических возможностей поиск, извлечение и отбор знаний из различных информационных ресурсов.

Современные технологии позволили обеспечить оцифровку архивов, книжных хранилищ, медиапродукции. Новые издания и произведения в больших объемах уже сохраняются в цифровом виде. Например, Государственная публичная научно-техническая библиотека (ГПНТБ) России планомерно проводит оцифровку изданий из своего фонда для Национальной электронной библиотеки (НЭБ), поддерживает информационную систему доступа к электронным каталогам библиотек сферы образования, науки и культуры.

Разнообразных хранилищ данных, банков знаний и информационных ресурсов становится всё больше, но при этом отсутствует единое поле для их взаимодействия и формирования национального инструментария управления знаниями. А это, как уже было сказано, вопрос не только полезности, удобства поиска и использования накопленных знаний, но и задача национальной безопасности страны. Тем более, что всем известные и широко используемые международные средства поиска и управления знаниями, такие как классификаторы, идентификаторы и т.п., далеко не во всем соответствуют нашим национальным рубрикаторам, да и не всегда удовлетворяют не только наши потребности, но и современное состояние и развитие отрасли контента.

Например, Всероссийский институт научной и технической информации (ВИНИТИ) РАН много лет ведет работу по поддержанию русскоязычного варианта международной Универсальной десятичной классификации. Однако последнее время Институт не всегда находит понимание у иностранных членов Консорциума УДК, предлагая модифицировать таблицы УДК в соответствии с изменениями, происходящими в современной науке и технике. Такое непонимание не отвечает национальным интересам России и зачастую просто мешает качественному выполнению исследовательской деятельности. Или, например, DOI (digital object identifier) – авторитетный и широко используемый в цифровом пространстве инструмент идентификации объектов, в том числе текстовых. Но многие страны из соображений национальной независимости и безопасности наряду с DOI вводят свои национальные идентификаторы. Россию в международном консорциуме *International DOI Foundation* представляет ВИНИТИ РАН. При этом в России есть свой идентификационный индекс – это национальный регистрационный номер издания, который выдает РКП. Она же выполняет функции национального агентства ISBN (международный уникальный номер книжного издания, необходимый для распространения книги в торговых сетях и автоматизации работы с изданием) и ISSN (международный уникальный номер, позволяющий идентифицировать любое периодическое издание независимо от того, где, на каком языке оно издано и на каком носителе размещено).

Вполне естественным является стремление РКП и ВИНИТИ РАН совместно разработать синтетические инструменты и технологии, приводящие в соответствие значения национальных и международных индексов.

Итак, в России есть своя национальная государственная регистрация изданий, развиваются международные, национальные и отраслевые классификаторы – ГРНТИ, Рубрикатор ВИНИТИ, ББК, УДК, действуют системы международной идентификации – ISSN, ISBN, DOI. Система обязательного экземпляра и национальная подписка обеспечивают доступ ко все-

возможным изданиям и информационным ресурсам. Очевидно, что всё это национальное достояние может жить и развиваться в рамках единой доступной среды. Такая открытая отечественная система цифровизации знаний, построенная на принципах взаимодействия информационных подразделений государственных ведомств – науки, культуры, образования, связи и массовых коммуникаций с учреждениями и организациями различных форм собственности страны, желающими обогатить свои возможности, пользуясь доступным контентом, отвечающим современным требованиям и учитывающим национальную специфику инструментарием, позволит в перспективе выстроить общество знаний с развитой информационной инфраструктурой.

Надо заметить, что концепция информационного кластера, включающего различные типы организаций страны, работающих с разнородной информацией любой отраслевой и ведомственной направленности, сегодня находит широкую поддержку профессионалов. В стране уже известны успешные примеры создания социально и профессионально ориентированных объединений, таких, например, как Гильдия книжников или Национальная библиотечная ассоциация «Библиотеки будущего». Подчеркнём, что сейчас ставится вопрос о государственной инфраструктуре знаний, формирующейся в рамках государственно-частного партнерства со всеми участниками индустрии знаний, объединенными не только общими интересами, но и образовательными и научными связями, общими историческими и культурными традициями. Причем в такой постановке вопроса речь, вероятно, должна идти и об организациях из СНГ и ЕАЭС.

Главное, что должна обеспечить инфраструктура знаний:

1) принципиально более высокий уровень доступа к знаниям для потребителя, а именно:

- территориальную доступность знаний, отсутствие ограничений по географическому принципу,
- техническую доступность знаний в любой форме издания – электронной или бумажной,
- оперативный доступ к актуальной информации, при котором пользователь должен получать доступ ко вновь опубликованному материалу в качественно более сжатые сроки,
- экономическую доступность, причем уровень издержек на потребление знаний должен быть приемлем для подавляющей части населения;

2) принципиально более быстрый и удобный поиск знаний для потребителя, а именно:

- бесплатный доступ к удобному единому универсальному электронному каталогу изданий,
- технологии, позволяющие осуществлять интеллектуальный контекстный и семантический поиск,
- четкую, непротиворечивую систему охраны прав на интеллектуальную собственность на всем её жизненном цикле.

В идее формирования инфраструктуры знаний нет ничего революционного. Как часто бывает, новое – это хорошо забытое старое. Инфраструктура научно-технических знаний существовала в СССР. Она называлась Государственная система научно-технической информации (ГСНТИ). Её базовыми структурными элементами тогда были Всесоюзный институт научной и технической информации АН СССР как головная организация, а также Государственная публичная научно-техническая библиотека, Всесоюзная книжная палата, Всесоюзный институт межотраслевой информации и некоторые другие.

Но время и технологии не стоят на месте, да и процесс конвергенции разнообразных знаний закономерно ведет к формированию универсальной инфраструктуры для всей индустрии знаний. И когда сегодня мы говорим про инфраструктуру знаний, то фактически ведём речь о своеобразной реинкарнации ГСНТИ практически в том же составе основных участников только на следующем витке развития общества.

Материал поступил в редакцию 26.09.17.

Сведения об авторах

БИКТИМИРОВ Марат Рамилевич – кандидат технических наук, ВРИО директора Всероссийского института научной и технической информации РАН
e-mail: marat@ras.ru

ЕСЕНЬКИН Борис Семёнович – доктор экономических наук, кандидат философских наук, заслуженный работник культуры РФ, президент НП «Гильдия книжников»
e-mail: esenkin@biblio-globus.ru

ЗОТОВ Павел Алексеевич – советник генерального директора ИТАР-ТАСС, исполнительный директор Российской книжной палаты
e-mail: zotov_p@tass.ru

НОГИНА Елена Борисовна – кандидат химических наук, директор Российской книжной палаты
e-mail: enogina@bookchamber.ru

ШРАЙБЕРГ Яков Леонидович – доктор технических наук, профессор, заслуженный работник культуры РФ, генеральный директор Государственной публичной научно-технической библиотеки
e-mail: shra@gpntb.ru

УДК 001.8

Р.С. Гиляревский

О научных публикациях, содержащих численные данные экспериментальных исследований*

Обосновывается предложение при разработке наукометрических систем вводить в библиометрические методы управления научными исследованиями показатель, учитывающий статьи с данными экспериментальных исследований. Только статьи этого типа можно однозначно соотнести с проведенным исследованием, тогда как растущее количество публикаций все меньше отражает реальный объем научных достижений.

Ключевые слова: *экспериментальные исследования, численные данные, управление научными исследованиями, наукометрические системы, библиометрические системы*

Как было предсказано в 1960-е гг., сегодня научные исследования перешли в стадию «большой науки», стали проводиться в крупных коллективах и потребовали значительных ассигнований из государственных и частных бюджетов. Перед инвесторами встал вопрос о контроле разумного использования потраченных учеными средств. Возникла также проблема собственности на результаты исследований. Эта проблема оказалась трудно разрешимой. Правительства и предприниматели, тратящие свои ресурсы на научные исследования, стали запрещать ученым открыто публиковать их результаты. Этой проблеме были посвящены дискуссии на международных [1] и национальных [2] конференциях. Стало ясно, что наука перестанет развиваться, если уничтожить систему научной коммуникации. Чиновники и бизнесмены поняли, что они не смогут оценивать эффективность исследований, особенно фундаментальных, если контроль не будут вести сами научные работники. И от запрещения публиковать результаты исследований пришлось отказаться.

Но ученые сами подсказали, как можно оценивать их труд, ничего не понимая в содержании проводимых ими исследований. В те же годы тогда еще молодой химик Ю. Гарфилд (1925–2017) создал особый информационный язык, основанный на библиографических ссылках. Он предложил искать нужные статьи не по фамилиям авторов, которые их написали, а по фамилиям тех авторов, на которых они ссылались, разумно предположив, что авторитеты больше известны [3]. Для этого он создал поисковую систему *Science Citation Index* (Указатель библиографических ссылок в естественных науках). Впоследствии она стала использоваться и для оценки научного труда. Публикационная активность ученого стала измеряться не только количеством его статей, но и количеством ссылок на эти статьи.

Следующий шаг был сделан уже в наше время. Объем научных достижений ученого, учреждения, в котором он служит, страны, в которой он работает, стал измеряться не количеством и качеством проведенных экспериментальных исследований, а публикационной активностью участников исследований. И вот уже на уровне национальных научных сообществ и государств ведутся сравнения этой активности, задаются ее плановые показатели, составляются отчеты и пишутся рапорты о приближении к этим показателям. Статистика, эта лукавая «игра в цыфирь» подменяет содержательную экспертную оценку научных достижений только потому, что в эту игру могут играть чиновники, не понимающие сути дела.

Мы сами повесили на себя этот хомут, теперь надо думать, как исправлять ситуацию. В подсчет библиографических ссылок вовлечены огромные силы и средства, этим заняты многотысячные коллективы, вооруженные самыми современными вычислительными инструментами и системами. Достаточно назвать таких гигантов, как фирмы США *Clarivate Analytics* (владеющая *Web of Science*) [4], Нидерландов *Elsevir* (создавшая *Scopus*) [5[, Германии *Springer*, России *eLibrary* (поддерживающая *РИНЦ*) и многие другие почти во всех развитых странах. В некоторых из них журнальные статьи и ссылки на них учитываются по их типам, в *Web of Science*, например, по двадцати разным типам (*article, meeting abstract, editorial material, review*). Но никто и нигде не выделяет

* Работа выполнена в соответствии с государственным заказом по теме 0003-2015-0008 «Разработка принципов государственной наукометрической системы» и в рамках проекта РФФИ «Исследование и разработка новых направлений использования наукометрических методов в прогнозировании, управлении научными исследованиями, оценке и стимулировании», грант №17-07-00256

в отдельный вид *статьи, содержащие численные данные экспериментальных исследований*, т.е. статьи, которые можно однозначно (один в один) соотнести с проведенным исследованием.

В практике научных публикаций обычно одному исследованию сопутствует несколько статей. Сначала публикуется заявочная статья, в которой закрепляется приоритет на идею и тему исследования, затем методическая, в которой излагаются его методы, дальше приводятся результаты исследования (т.е. численные данные эксперимента), иногда отдельно публикуется обсуждение результатов с указанием областей их применения. Поскольку в последние годы во всех странах активно стимулируется публикационная активность научных работников, количество журнальных статей быстро опережает количество проводимых исследований. В общей массе учитываемых статей все больше превалируют обзоры, дискуссии, рассуждения, терминологические и науковедческие исследования.

В последние годы усилился поток статей по наукометрии и так называемому цитированию. Разумеется, теоретические, дискуссионные, терминологические и обзорные статьи отражают проведенные исследования и важны для развития науки. Но данные, полученные в результате экспериментов, служат основой этого развития. Таким образом, количество публикаций все меньше отражает реальный объем научных достижений. Это предположение основано на профессиональной интуиции многолетнего редактирования журнала «Научно-техническая информация», но оно нуждается в проверке.

Чтобы не подвергаться искушению манипулировать статистическими данными, вернее всего было бы *de visu* посмотреть несколько сотен журналов за десятилетие и выявить динамику в соотношении публикаций, содержащих численные данные экспериментальных исследований, с общим потоком публикаций всех типов. Но для этого даже в ста журналах за десять лет пришлось бы просмотреть более ста тысяч статей, что нереально. Поэтому для проверки высказанного предположения пришлось воспользоваться доступными массивами вторичной информации и методами автоматического распознавания таких статей. Что касается массивов вторичных данных, то ими послужили банк данных ВИНИТИ РАН, из которого мы выгрузили рефераты за 2006–2015 гг., опубликованные в реферативных журналах по информатике и молекулярной биологии общим объемом 92 тыс. записей, а также данные *Web of Science* за те же годы объемом 1,7 млн записей.

Предварительно нами были просмотрены выпуски журналов «Научно-техническая информация» и «Международный форум по информации» за последнее десять лет, выявлены статьи интересующего нас типа и из них выбраны часто встречающиеся слова и словосочетания, формально указывающие на принадлежность к этому типу. Динамика изменения доли этих статей в этих журналах наших предположений не подтвердила.

Затем выбранные слова и словосочетания были автоматически сверены с рефератами в названных выше базах данных ВИНИТИ. Слова также не дали никаких результатов, кроме слишком частой встречаемости некоторых из них в рефератах статей разных типов (например, *анализ** 28 тыс., *данны** 18 тыс., *результат** 12 тыс. раз). Результаты проверки одного из словосочетаний показаны в табл. 1.

Поисковый запрос на словосочетание *анализ данных* со всей парадигмой изменений по падежам и числам (усечения слов показаны звездочкой) свидетельствует о некотором снижении доли статей, содержащих численные данные экспериментальных исследований в обоих массивах рефератов. Анализируемые реферативные журналы были выбраны по контрасту, так как в информатике нужных нам типов статей немного, а молекулярная биология преимущественно экспериментальная область науки. Однако полученные результаты не были убедительны по ряду причин. Далеко не все выбранные нами словосочетания (например, *результаты исследования, данные эксперимента, разработка методов*), характерные для данного типа изданий, показали снижение доли статей. Все эти словосочетания встречаются и в других типах статей, хотя и с меньшей частотностью. Очень велик разброс данных по годам.

Проверка полученных данных на более крупных массивах БД *Web of Science* по аналогичным поисковым запросам на английском языке дала те же результаты (табл. 2) и породила те же сомнения в их убедительности.

По всей вероятности, попытка выйти на определенный тип статьи путем анализа лексики требует более сложных методов и систем семантического анализа, нежели простая статистика.

Таблица 1

Динамика изменения доли публикаций (%), содержащих результаты исследований, по информатике и молекулярной биологии в РЖ ВИНИТИ РАН в общем числе публикаций этого раздела в определенном году

Реферативный журнал	Поисковый запрос	2006	2015
Информатика	*анализ* AND данны**	5,3	5,2
	таблицы	17,6	10,2
	обзор	3,7	5,3
Молекулярная биология	*анализ* AND данны**	16,3	9,2
	таблицы	13,5	10,7
	обзор	3,9	9,0

Таблица 2

Динамика изменения доли публикаций (%), содержащих результаты исследований, в БД *Web of Science* в разделах *Information Science & Library Science* и *Biochemistry & Molecular Biology* в общем числе публикаций этого раздела в определенном году

Название БД	Поисковый запрос	2006	2015
Information Science & Library Science	*analysis* AND data**	0,09	0,05
	result AND experiment**	0,02	0,01
	review	0,8	1,2
Biochemistry & Molecular Biology	*analysis* AND data**	7,9	6,9
	result AND experiment**	4,7	5,1
	review	6,8	8,3

Поскольку ни в одном массиве первичных или вторичных документов интересующий нас тип документов не учитывается отдельно (документы типа *data* в *Web of Science* не принимаются в расчет, так как введены год назад и имеют несколько иное назначение), были предприняты поиски формальных элементов, указывающих на принадлежность статей к данному типу. Таким элементом могли бы стать таблицы, поскольку именно они используются для показа численных данных. К сожалению, в доступных массивах вторичной информации наличие таблиц указывается редко и нерегулярно. Там, где было возможно, в табл. 1 введены сведения о динамике этого элемента в статьях. Не очень полные сведения имеют отрицательную динамику, свидетельствующую как бы о снижении доли статей с числовыми данными экспериментов. Но и они не очень убедительны, так как таблицы могут содержаться и в других типах статей.

В ходе дальнейших поисков доказательства правильности высказанной гипотезы о снижении доли статей с результатами экспериментальных исследований было обращено внимание на поисковый запрос *обзор* в табл. 1 и на аналогичный запрос *review* в табл. 2. Они показывали устойчивую положительную динамику и значительно меньший разброс по годам. Это навело на мысль попытаться пойти от обратного и использовать статистику обзорных статей, которые заведомо не могут быть результатом экспериментального исследования. Если динамика увеличения их доли в общем потоке публикаций значительно превышает динамику общего увеличения количества статей, то это может быть косвенным свидетельством стабильности или даже уменьшения количества проводимых экспериментов.

Имеющиеся в БД *Web of Science* сведения о типах статей позволяют провести такое исследование. Чтобы быть последовательными, мы не меняли тематики, но расширили временные границы для лучшего понимания тенденции в динамике изменения доли обзоров в разделе информационно-библиотечной науки, соответствующей содержанию РЖ Информатика в *Web of Science* (на рис. 1), а на рис. 2 показано то же по биохимии и микробиологии. Наибольший интерес представляет рис. 3, на котором представлена динамика изменения доли обзоров в БД *Web of Science* за прошедшие сто семнадцать лет.

Большой разброс доли обзоров по годам и отсутствие внятной тенденции ее изменения в публикациях по информационно-библиотечной науке на графике (см. рис. 1) не дает уверенности в наших предположениях об увеличении в последние десятилетия доли статей, не содержащих числовых данных экспериментальных исследований. Но два других графика (см. рис.2 и рис. 3) полностью подтверждают это предположение. Доля обзоров по биохимии в трех основных базах данных *Web of Science* (*Science Citation Index_Social Sciences CI, Art &Humanity CI*) в целом демонстрирует резкое их увеличение в общей массе публикаций, в десятки раз в первом случае и почти в семь раз во втором.

В структуре всех семи баз данных *Web of Science* обзоры составляли в 1975 г. 1,1% и в 2016 г. 4,3%. За этот период абсолютное количество обзоров увеличилось с 6 тыс. до 92 тыс. т.е. более чем в 15 раз, тогда как общий поток публикаций в этой БД возрос с 566 тыс. до 2,1 млн., т.е. всего лишь в 3,7 раза. При этом количество публикаций, не являющихся собственно научными статьями (материалы конференций, письма и редакционные статьи), за это время оставалось относительно стабильным и даже снизило свою долю в общем потоке с 20% до 16,7%. Это свидетельствует о том, что с середины 1970-х гг., т.е. с начала активной кампании по стимулированию публикационной активности доля научных статей, в том числе содержащих экспериментальные данные, снижается в общем потоке публикаций.

Если верно то, что публикация данных – это признак проведения экспериментального исследования и что именно это исследование является основой развития и достижений науки, то придется признать несостоятельность нынешних библиометрических данных о росте количества публикаций в качестве показателей этого развития и достижений. А ведь именно на них во многом основывается сравнение научного потенциала ученых, организаций и даже стран.

Рост количества публикаций ученого, организации или даже научного сообщества страны вовсе не свидетельствует об их научных достижениях, о развитии науки в целом. Этот рост указывает лишь на успех искусственного стимулирования так называемой публикационной активности, которая дезорганизует научную деятельность.

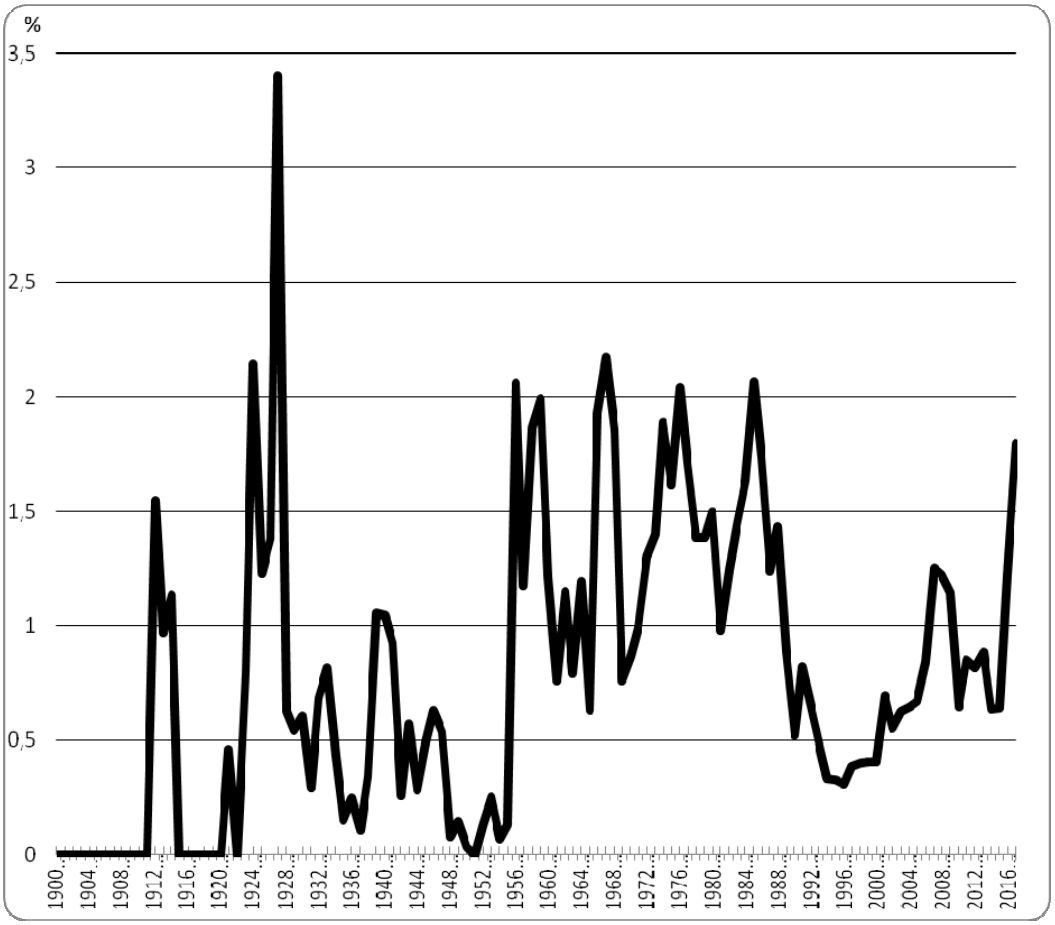


Рис. 1. Динамика изменения доли обзоров (%) в БД *Web of Science* (*SCI_SSCI_A&HCI*) в разделе *Information Science & Library Science* в общем числе публикаций этого раздела в определенном году

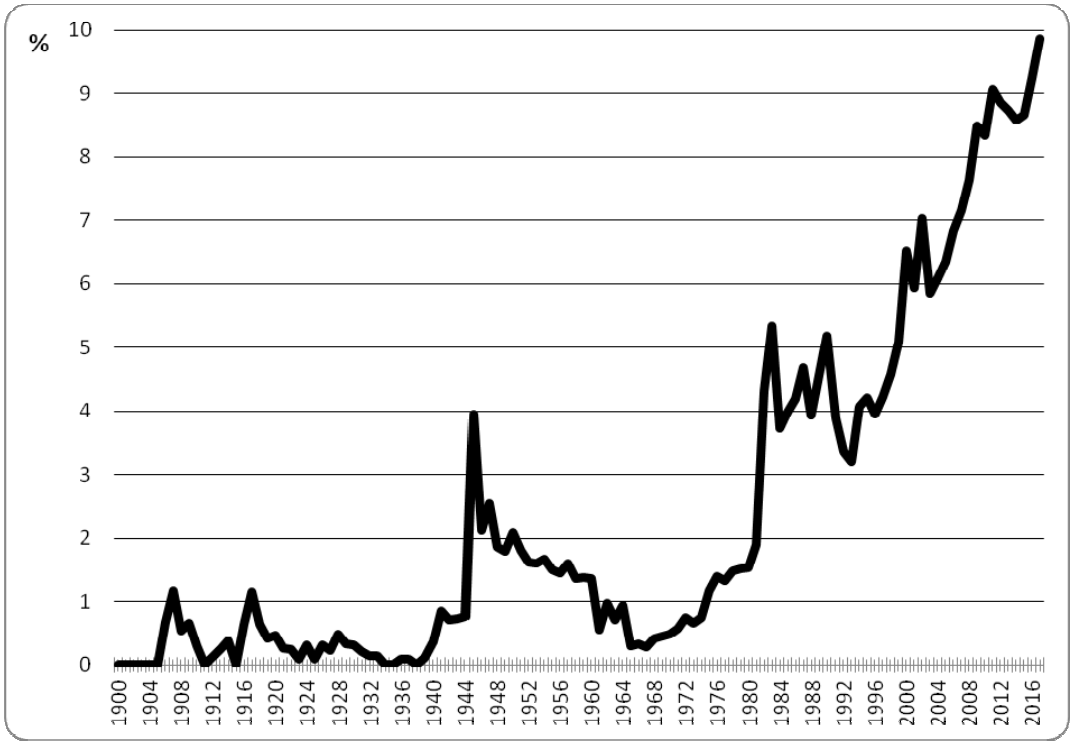


Рис. 2. Динамика изменения доли обзоров (%) в БД *Web of Science* (*SCI_SSCI_A&HCI*) в разделе *Biochemistry & Molecular Biology* в общем числе публикаций этого раздела в определенном году

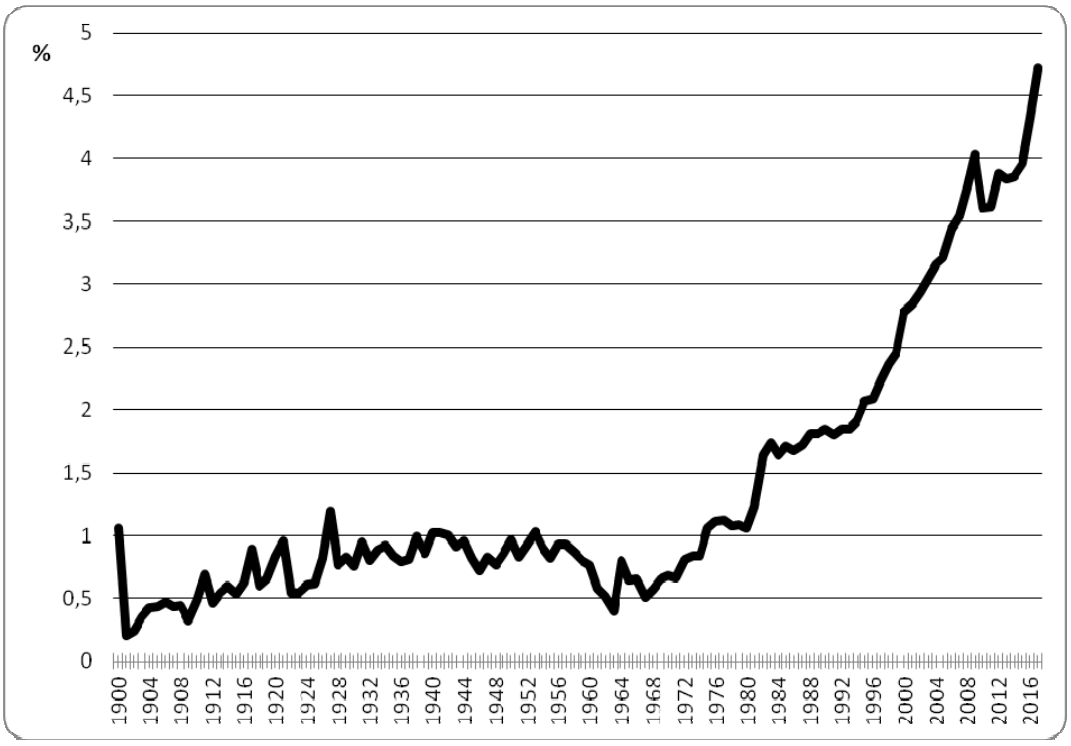


Рис. 3. Динамика изменения доли обзоров (%) в БД *Web of Science* (*SCI_SSCI_A&HCI*) в общем числе публикаций в определенном году

Большинство ученых, особенно в точных и естественных науках, хорошо понимают это без специальных подсчетов. Президент Российской академии наук, академик А.М. Сергеев недавно сказал в газетном интервью: «Что касается числа публикаций в престижных изданиях, то есть мнение, что мы растем. Но на самом деле этот рост искусственный. Мы являемся соавторами в длинном списке по результатам, полученным на зарубежных установках. А вот публикаций высокого уровня на основе экспериментов, сделанных в России, совсем мало» [6]. Однако люди, управляющие наукой, а «большая наука» нуждается в управлении менеджерами, этого понимать не хотят, чтобы не лишится такого якобы объективного инструмента управления.

Напрашивается следующий вывод. В национальной наукометрической системе при разработке библиометрических методов управления научными исследованиями необходимо предусмотреть показатель, учитывающий статьи с данными экспериментальных исследований.

Любые численные показатели в библио-, инфор-, наукометрии и других «метриях» нуждаются интерпретации и осторожном использовании, особенно при оценке интеллектуального труда. В данном случае желательно сопоставлять эти данные с экспертной оценкой. «Ученые, несомненно, нуждаются в оценке их труда, которая является важным стимулом их деятельности, – писал биолог Ганс Селье, открывший стресс. – Ни один ученый, достойный этого звания, не измеряет свой успех количеством похваливших его людей,… ученые нуждаются в периодическом одобрительном "похлопывании по плечу", точно так же, как и все простые смертные, хотя они, по тем или иным причинам, не очень-то склонны в этом сознаваться. Разумеется, значение имеет не шумовой уровень аплодисментов, а кто и за что вам аплодирует» [7].

Руководство американской фирмы *Bell Telephon Laboratories* прекратило как несоответствующие тематике лаборатории исследования Ч. Таунса [8], за которые он вместе с А.М. Прохоровым и Н.Г. Басовым получил Нобелевскую премию за основополагающие работы по квантовой электронике. Известно, что Дж. Уотсон [9], Ф. Крик и М. Уилкинс получили Нобелевскую премию за выяснение структуры молекулы ДНК, опубликовав в журнале «*Nature*» одну статью в девятьсот слов. Считаю, что любой, кто претендует руководить наукой как социальным явлением, будь он заведующим лабораторией, президентом академии наук или министром, должен прочитать хотя бы упомянутые книги. И еще многие другие. В их числе: Кун Т. Структура научных революций. – М.: Прогресс, 1977; Прайс Дж.Д. Малая наука, большая наука // Наука о науке. – М.: Прогресс, 1966. – С. 281–384.

* * *

Автор выражает благодарность ведущему научному сотруднику ВИНИТИ РАН Александру Наумовичу Либкинду за большую помощь в обработке и интерпретации статистических данных.

СПИСОК ЛИТЕРАТУРЫ

1. UNESCO symposium, Paris, March 10–11, 2003.– URL https://www.nap.edu/catalog/
2. Open access and the public domain in digital data and information for science: Proceedings of an international symposium. – Washington: National Academies Press, 2004.
3. Garfield E. Science Citation Index: a new dimension in indexing // Science. – 1964. – Vol.144, № 96. – P.649-654.
4. Гиляревский Р.С., Мельникова Е.В. Институт научной информации США: Идеология, преобразования, продукты // Научно-техническая информация. Сер.1.– 2017.– № 10. – С. 26-31.
5. Мельникова Е.В. Издательство *Elsevier* и информационная система *Scopus* // Научно-техническая информация. Сер.1. – 2017. – № 7. – С. 19-22.
6. Наука Академии // Российская газета. – 2017. – 9 октября. – С. 7.
7. Селье Г. От мечты к открытию : Как стать ученым / пер. с англ. – М.: Прогресс, 1987.
8. Таунс Ч. Квантовая электроника и технический прогресс: Проблема планирования исследовательской работы // Успехи физических наук. – 1969. – Т. 98, вып.1. – С. 159–169.
9. Уотсон Дж. Д. Двойная спираль. – М.: AST, 2013.

Материал поступил в редакцию 26.09.17.

Сведения об авторе

ГИЛЯРЕВСКИЙ Руджеро Сергеевич – доктор филологических наук, профессор, заведующий отделением научных исследований по проблемам информатики ВИНИТИ РАН; профессор факультета журналистики Московского государственного университета им. М.В. Ломоносова
e-mail: giliarevski@viniti.ru

ОРГАНИЗАЦИЯ ИНФОРМАЦИОННОЙ РАБОТЫ

УДК 002.6 – 047.44

Д.А. Девяткин, Р.Е. Суворов, И.А. Тихомиров

О методике выявления центров компетенции на примере предметной области «Искусственный интеллект»*

Предлагается методика выявления центров компетенции с использованием больших коллекций научно-технических документов и методов компьютерной лингвистики. Для апробации методики в качестве примера предметной области выбран искусственный интеллект. Исследования проведены на основе полных текстов трудов национальной конференции по искусственному интеллекту, авторефератов кандидатских и докторских диссертаций и данных РИНЦ. Отобраны основные специальности, по которым защищаются диссертации по искусственному интеллекту, построена динамика защит, получены наукометрические показатели ученых, работающих в области искусственного интеллекта, выявлены основные центры компетенции и их специализация.

Ключевые слова: *анализ текстов, методика выявления центров компетенции, центры компетенции, искусственный интеллект, публикационная активность, рейтинг организаций, РИНЦ*

ВВЕДЕНИЕ

Рано или поздно для любой предметной области возникает задача выявления центров компетенции. Очевидным решением этой задачи могло бы быть использование таких баз данных, как *Scopus*, *Web of Science* или РИНЦ. К сожалению, на данный момент РИНЦ не может быть достоверным источником информации для решения поставленной задачи: имеются проблемы с точностью соотнесения публикаций и авторов, данные не всегда имеют должное качество, может присутствовать «накрутка» показателей [1–3] и т.д. Качество зарубежных баз цитирования заметно выше, однако для некоторых традиционно сильных российских научных направлений в них отражается мало публикаций, что затрудняет или делает невозможным использование этих баз данных для выявления центров компетенции по целому ряду предметных областей.

При решении задачи выявления центров компетенции возникает проблема идентификации анализируемой предметной области. Для этой цели обычно используют системы классификации: ГРНТИ, УДК, Высшей аттестационной комиссии (ВАК), МПК, РИНЦ, *Web of Science, Scopus* и др. К сожалению, коды классификаторов покрывают далеко не все существующие направления исследований. Еще одной проблемой является то, что эти классификаторы не связаны между собой и сильно отличаются друг от друга, что затрудняет интеграцию данных из различных источников. Кроме того, изменение классификаторов со временем приводит к исчезновению одних кодов и появлению новых. Бывает и так, что анализируемая предметная область далеко не всегда легко укладывается в таксономию, предлагаемую авторами классификаторов. В результате привязка конкретных исследований к кодам бывает весьма спорна, а зачастую и некорректна.

Таким образом, задача выявления центров компетенции может оказаться совсем непростой ввиду возможного отсутствия соответствующих направлению кодов классификаторов (тематических таксономий), их неполноты и неточности, а также целого ряда проблем с качеством и полнотой отечественных и зарубежных баз цитирования.

Цель настоящего исследования – разработка и апробация методики выявления центров компетенции на основе данных о диссертационной и публикационной активности ученых России с применением средств автоматического анализа полных текстов.

ХАРАКТЕРИСТИКА ПРЕДМЕТНОЙ ОБЛАСТИ ДЛЯ АПРОБАЦИИ МЕТОДИКИ

Методика выявления центров компетенции, представленная в настоящей работе, апробирована на устоявшейся научной дисциплине – искусственный интеллект (ИИ). Если говорить просто, то основная

цель ИИ – это решение сложных задач, с которыми успешно справляется человек, но для которых неизвестен простой алгоритм решения, пригодного для выполнения на современных вычислительных машинах [4]. Как правило, отличительная черта заключается в том, что алгоритм решения задачи является одним из результатов её решения. К подобластям искусственного интеллекта относят приобретение знаний, машинное обучение, анализ естественного языка, интеллектуальное планирование и многое другое.

Несмотря на то, что искусственный интеллект сформировался как отдельная научная дисциплина достаточно давно, в номенклатуре ВАК нет научной специальности, соответствующей именно искусственному интеллекту. Специалисты по ИИ защищаются по таким специальностям, как «Системный анализ, управление и обработка информации», «Теоретические основы информатики» и другим. Если рассмотреть существующие классификаторы, то в УДК для ИИ отведен специальный код 004.8, в ГРНТИ 28.23, а в базе данных *Scopus* – 1702. В тематическом рубрикаторе РИНЦ специальный код для ИИ отсутствует, однако имеются близкие по тематике 20.00.00 «Информатика» и 28.00.00 «Кибернетика».

В Scopus за 2006-2015 гг. проиндексировано всего 1367 работ российских ученых по ИИ, тогда как поиск в РИНЦ только по одной связанной тематике «Кибернетика» за аналогичный период дает более 12,1 тыс. публикаций российских авторов.

Резюмируя изложенное можно утверждать, что работы российских ученых по искусственному интеллекту слабо представлены в зарубежных базах цитирования, коды классификаторов, соответствующих ИИ рознятся, а в некоторых классификаторах так и вовсе отсутствуют.

МЕТОДИКА ВЫЯВЛЕНИЯ ЦЕНТРОВ КОМПЕТЕНЦИИ

Первый шаг, предлагаемой нами методики – идентификация предметной области. Для этого вручную формируется некоторый начальный набор текстов, посвященных анализируемой предметной области, а затем с помощью метода [5] выполняется поиск авторефератов диссертаций, похожих на эти тексты. Указанный метод специально разработан для вычислительно эффективного решения задачи поиска тематически похожих документов с применением инвертированных индексов документов [6]. Для каждой статьи из начального набора отдельно ведется поиск похожих авторефератов, а результаты поиска объединяются в один список, при этом гиперпараметры метода поиска похожих документов подбираются эмпирически. Для актуализации результатов поиска рассматриваются только работы, выполненные в течение 10 лет до момента исследования. В России оценка научной квалификации научных работников обеспечивается государственной системой научной аттестации, которая предусматривает присуждение ученых степеней в соответствии с номенклатурой научных специальностей. Следовательно, можно утверждать, что исследователь, защитивший диссертацию по соответствующей отрасли науки, обладает определенным уровнем знаний в рамках тематики его диссертационного исследования, что, как ожидается, должно повысить качество выявления центров компетенции. С использованием метаданных найденных авторефератов выявляется топ специальностей, по которым производятся защиты в анализируемой предметной области, и строится динамика защит. Для повышения точности анализа авторефераты диссертаций по специальностям, не вошедшим в топ, отфильтровываются.

На втором шаге методики сопоставляются найденные авторефераты и профили исследователей в РИНЦ, откуда загружаются наукометрические показатели. В рамках методики вводится допущение о том, что в РИНЦ консолидировано хранятся наиболее точные данные о текущих местах работы и наукометрии ученых, защитивших диссертации в последние 10 лет. Наукометрические показатели собираются в автоматизированном режиме. Поиск выполняется в авторском указателе *eLIBRARY* по ФИО ученого, затем запускается программный модуль, копирующий необходимые наукометрические показатели и информацию о месте работы в базу данных. В ходе дальнейшего анализа учитываются только те ученые, у которых за последние 5 лет была проиндексирована хотя бы одна публикация в РИНЦ. Таких ученых далее будем называть «активными».

На третьем шаге формируются два рейтинга центров компетенции по количеству исследователей, работающих в исследуемой научной области. Первый рейтинг строится с учетом всех активных ученых. При построении второго рейтинга предлагается учитывать только специалистов, чьи наукометрические показатели (индекс Хирша без учета самоцитирований и средний импакт-фактор журналов, в которых опубликованы результаты исследований) превышают медианные. Для выявления специализации организаций, вошедших в рейтинг, по текстам авторефератов сотрудников с помощью метода [7] строится ключевая лексика.

При необходимости выполняется *четвертый шаг*, на котором производится сравнительный анализ центров компетенции. В зависимости от целей исследования, центры компетенции могут разбиваться на группы. В ходе сравнительного анализа сопоставляются средние наукометрические показатели активных ученых, работающих в центрах компетенции, тематика исследований и т.д.

ИСХОДНЫЕ ДАННЫЕ ДЛЯ ПРОВЕДЕНИЯ ИССЛЕДОВАНИЙ

В качестве исходных данных для исследований нами использованы полные тексты докладов на конференциях по искусственному интеллекту КИИ-2014 и КИИ-2016, авторефераты кандидатских и докторских диссертаций, загруженные с сайта Российской государственной библиотеки (РГБ), а также наукометрические показатели РИНЦ (http://elibrary.ru).

Полные тексты докладов национальных конференций по искусственному интеллекту (КИИ) за 2014-2016 гг. используются для идентификации предметной области. Допущение заключается в том,

что большинство работ в области искусственного интеллекта в России должно в высокой степени отражать статьи, опубликованные в трудах этих конференций. Исходная выборка публикаций включает 238 статей.

Анализ публикационной активности ученых выполнялся на основе следующих наукометрических показателей из РИНЦ: индекс Хирша без учета самоцитирований; количество публикаций автора, отраженных в РИНЦ; средневзвешенные импакт-факторы журналов, в которых были опубликованы и процитированы статьи.

Эти показатели собирались в автоматизированном режиме. Выполнялся поиск в авторском указателе eLIBRARY по ФИО ученого, затем осуществлялся отбор по месту работы. Всего по заполненным профилям было найдено 4062 ученых, работающих в области искусственного интеллекта.

АПРОБАЦИЯ МЕТОДИКИ. РЕЗУЛЬТАТЫ ИДЕНТИФИКАЦИИ ПРЕДМЕТНОЙ ОБЛАСТИ «ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ»

Первая подзадача в рамках задачи выявления центров компетенции — сбор первичного материала. В данном случае в качестве первичных материалов выступают полные тексты авторефератов диссертаций, тематически относящиеся к статьям, опубликованным в рамках КИИ-2014 и КИИ-2016. В результате выявлены пять специальностей, по которым защищались найденные диссертации:

05.13.01 системный анализ, управление и обработка информации,

05.13.18 математическое моделирование, численные методы и комплексы программ,

05.13.06 автоматизация и управление технологическими процессами и производствами,

05.13.11. математическое обеспечение вычислительных машин, комплексов и компьютерных сетей,

05.13.17 теоретические основы информатики.

Динамика защит показана на рис. 1а. На 2013 г. пришелся пик количества защит по искусственному интеллекту, однако начиная с 2014 г. этот показатель снизился до уровня 2012 г. (рис. 1б).

Далее выявление центров компетенции производилось по активным ученым, защитившимся по одной из пяти перечисленных выше специальностей. Количество таких ученых составило 817.

АНАЛИЗ НАУКОМЕТРИЧЕСКИХ ПОКАЗАТЕЛЕЙ АКТИВНЫХ УЧЕНЫХ, РАБОТАЮЩИХ В ОБЛАСТИ ИИ

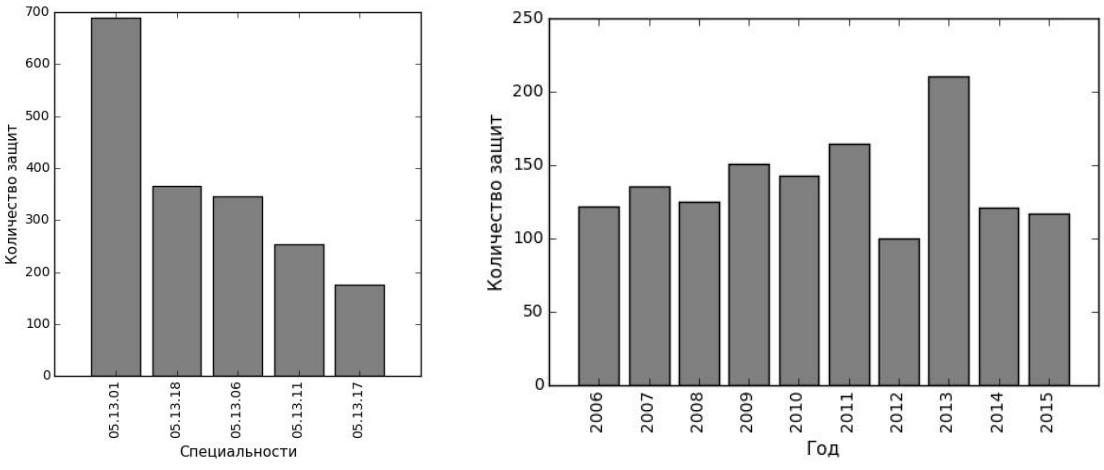
Из РИНЦ загружались и анализировались наукометрические показатели ученых, работающих в области искусственного интеллекта. В ходе анализа были исключены ученые, не опубликовавшие ни одной работы в период с 2010 по 2015 гг. На рис. 2 показано распределение ученых, работающих в России в области ИИ, по индексу Хирша: видно, что большая часть исследователей имеет индекс Хирша в диапазоне 1-6.

Если же не учитывать самоцитирования, то индекс Хирша большей части ученых находится в диапазоне 1-5 (рис. 3). Медиана этого показателя составляет 2. Медиана импакт-фактора журналов, в которых публикуются эти ученые, составляет 0,32 (рис. 4).

ЦЕНТРЫ КОМПЕТЕНЦИИ В ОБЛАСТИ ИИ

На основе информации об основном месте работы исследователей, полученной из профилей РИНЦ, был составлен список организаций, в которых работает наибольшее число активных ученых. Всего было сформировано 2 варианта списка: при построении первого учитывались все активные ученые, при составлении второго – только такие исследователи, чей индекс Хирша без самоцитирований превысил 2 и средний импакт-фактор журналов, в которых они публиковались, был выше 0,32 (т.е. с показателями выше медианных значений).

Полученные рейтинги представлены на рис 5-6. Необходимо отметить, что введение подобной фильтрации привело к тому, что значительное число научно-исследовательских учреждений РАН заняло более высокие позиции.



а) б)

Рис. 1а. Топ 5 специальностей и 1б. Динамика защит по искусственному интеллекту

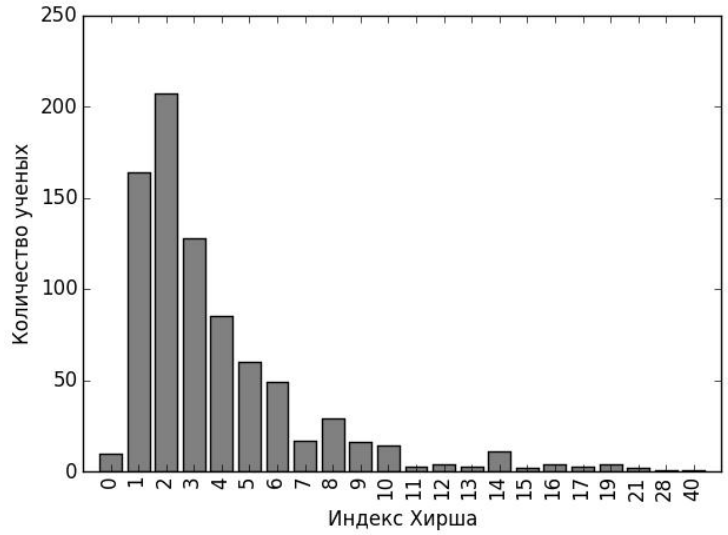


Рис. 2. Распределение ученых, работающих в области искусственного интеллекта, по индексу Хирша

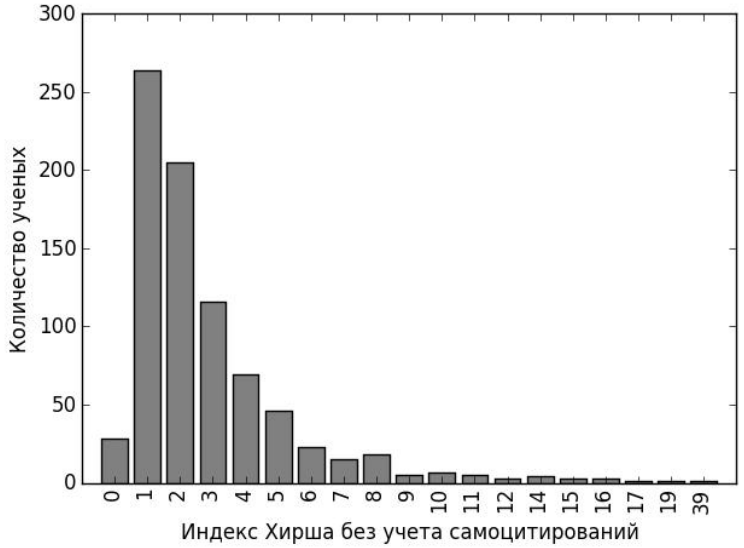


Рис. 3. Распределение ученых, работающих в области искусственного интеллекта, по индексу Хирша без учета самоцитирований

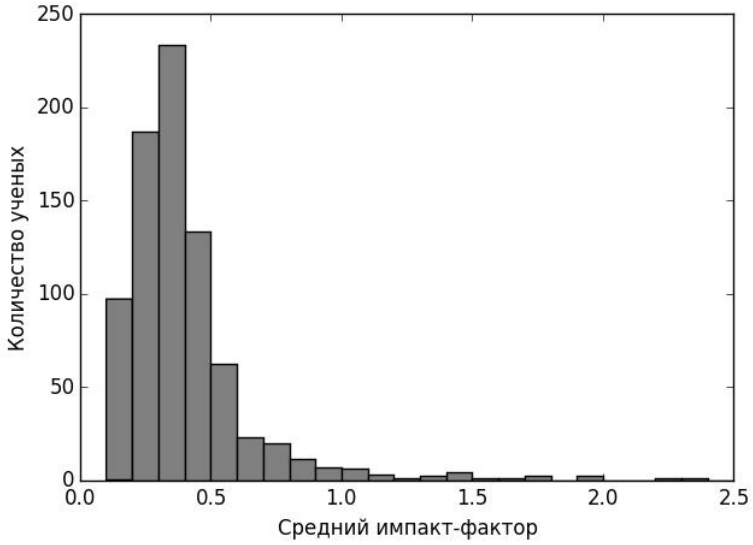


Рис. 4. Средний импакт-фактор научных журналов, в которых публикуются ученые, работающие в области искусственного интеллекта в России

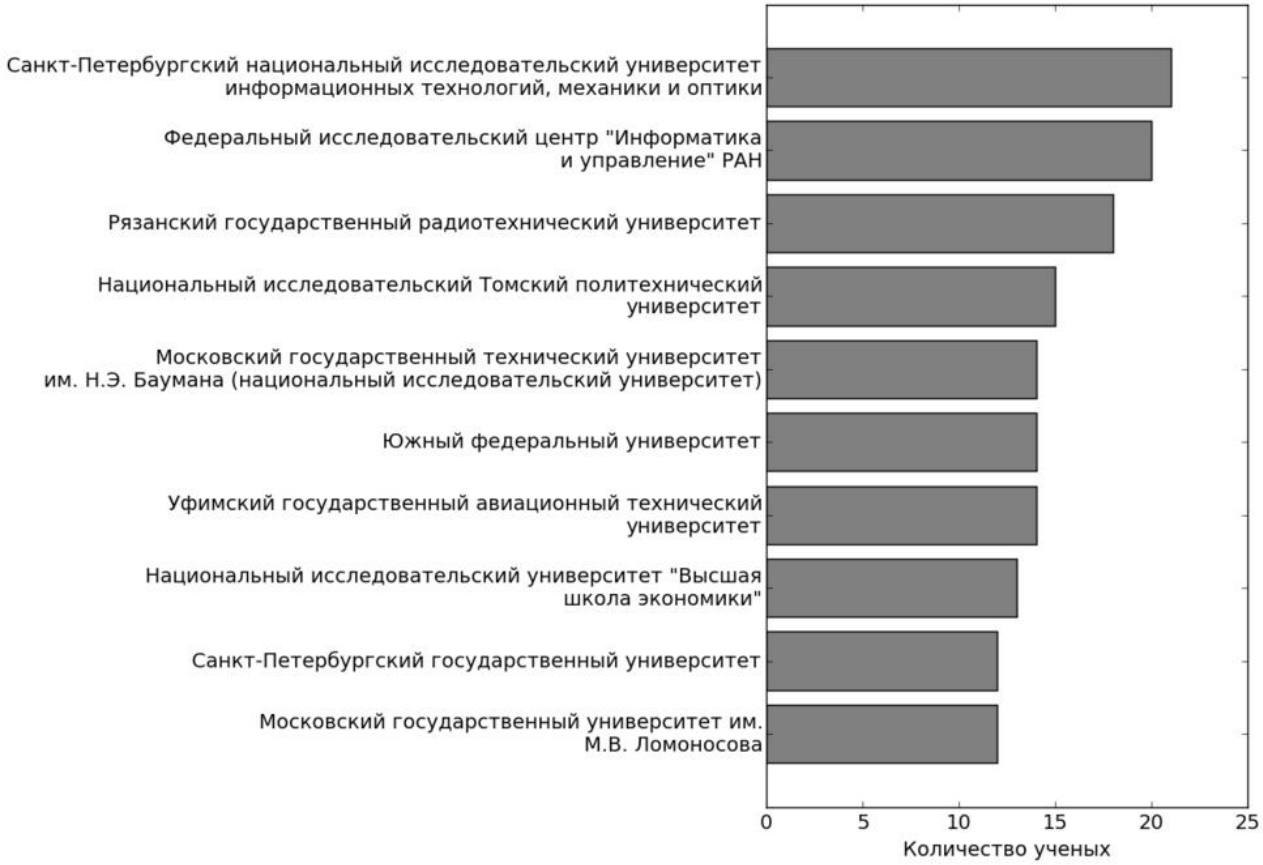


Рис. 5. Рейтинг организаций по количеству активных ученых в области искусственного интеллекта

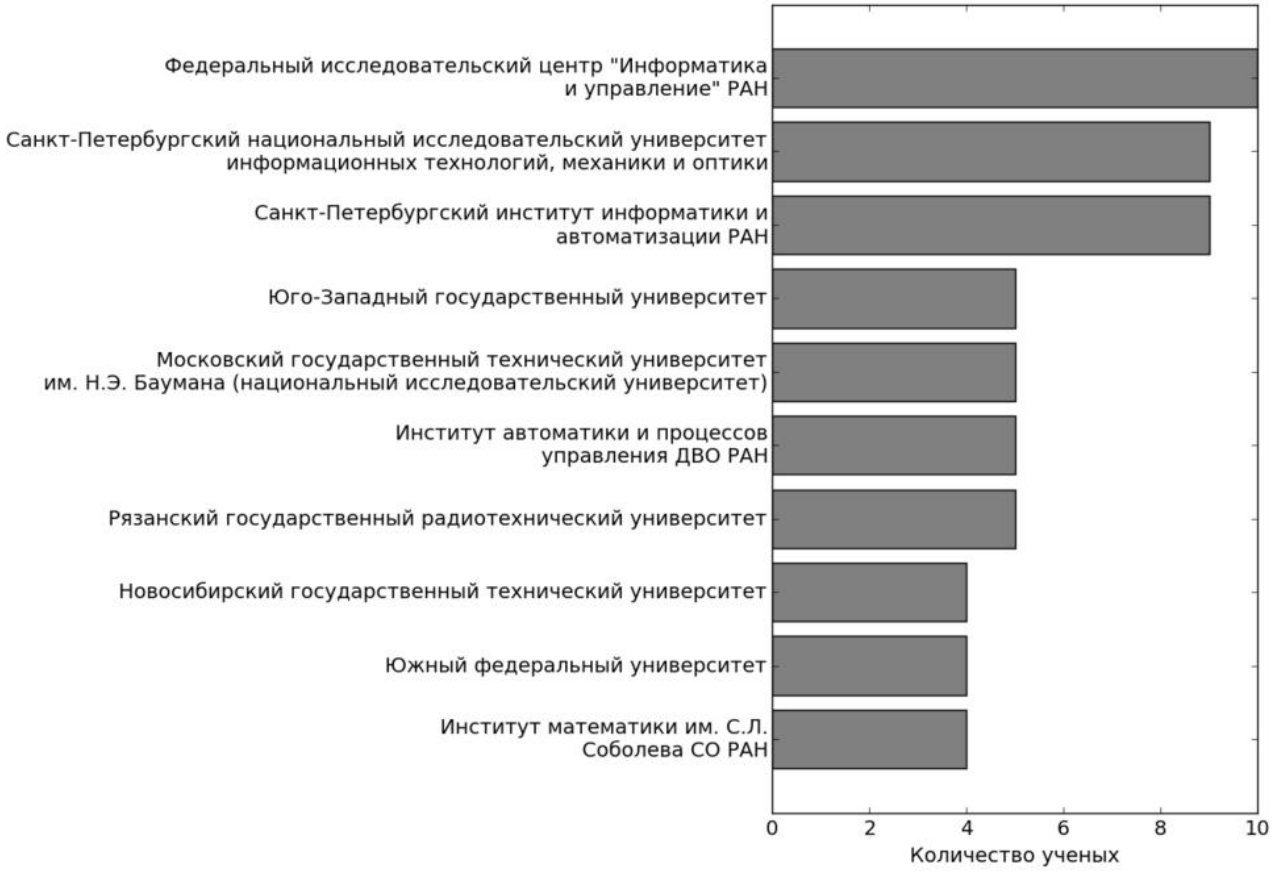


Рис. 6. Рейтинг организаций по количеству активных ученых в области искусственного интеллекта, отфильтрованных по наукометрическим показателям

Для выявления специализации организаций в области искусственного интеллекта по текстам авторефератов активных ученых была построена таблица ключевой лексики (таблица), из которой видно, что исследователи ФИЦ «Информатика и управление» РАН специализируются на задачах поиска и анализа текстов, в то время как специализация МГТУ им. Н.Э. Баумана – интеллектуальное управление.

Ключевая лексика, автоматически построенная по авторефератам диссертаций активных исследователей, работающих в организациях из отфильтрованного рейтинга

Организация	Ключевая лексика
Федеральный исследовательский центр "Информатика и управление" РАН	перевод, синтез гиис, грид, ролевая структура, корректность замыкания, критерий остановки, особенность онтологии, синтаксический, поисково-аналитическая обработка, кластеризация, язык-посредник, *intelligent method*, лексико-морфологическая информация, статический оператор, последовательность клеток, ролевой, ресурс-посредник, форма лексемы, глубина стратегии, абдуктивное принятие, модель-посредник, *hga* алгоритм, варьирование пространства, *descriptive image*
Санкт-Петербургский национальный исследовательский университет информационных технологий, механики и оптики	устранение цикличности, извлечение метаданных, общесистемный метод, микроклассификация, декомпозиция графов, встраивание пакета, качественная устойчивость, многопрофильная платформа, неизвестный результат, конформационно-зависимое свойство, сбалансированное обучение, нередуцированный граф, компьютерная симуляция, виртуальный полигон, извлечение онтологий, набор подалфавитов, экспериментальное опробование
Санкт-Петербургский институт информатики и автоматизации РАН	транскрипция словоформы, синтез шифра, уникальная транскрипция, речь, чтение, базовая транскрипция, жестовый язык, ассистивный, информационный резонанс, кумулятивный риск, модификация дерева, информативность, морфофонемный, лексическое дерево, представление словаря, матрично-векторное уравнение, аудиовизуальная модальность, команда на движение, префиксное дерево, фиксированная инволюция, шаблон обнаружения
Юго-Западный государственный университет	переразмещение программы, операционная часть, ускоренное планирование, макроэлемент изображения, генерировать альтернативу, высокоточная обработка, битовый признак, производительность, ограниченные затраты, двухкоординатная адресация, манжета механизма, актуализация модуля, повысить нагрузку, оперативно-тактическая обстановка, линейное программирование, замыкание алгоритма
Московский государственный технический университет им. Н.Э. Баумана (национальный исследовательский университет)	контрольный прецедент, управление, параметр, микроспутник студенческий, высокоточный электропривод, физический агент, малый срыв, внутрикластерный анализ, пересчет плана, детерминировать часть, нечетко-множественный, Пирсон, нейродиагностика, электродвигатель, модель обеспечивает, адекватный формализм, метрика управления
Институт автоматики и процессов управления ДВО РАН	специфика платформы, движитель, *advanced intelligent, automation mechatronics*, позиционно-силовое управление, контекстнозависимый, канал степень, интеллектуальный сервис, нечеткая логика, корректность остановки, отдельный движитель, нечеткое дерево, отрабатывать воздействие, агент, модельная сцена, калибровка изображения, параметр движителя, онтология, корреляционное сравнение, трифокальный, датчик нпа
Рязанский государственный радиотехнический университет	изоморфизм, алгоритм упорядочивания, Пирсон, одноуровневая кластеризация, суждение эксперта, стабилизация фона, объединение, элемент множества, кратномасштабная обработка, снижение восприятия, выявить патологию, норма, способ лечения, кластеризация, установление релевантности, множество центроид, переменная
Новосибирский государственный технический университет	объектно-ориентированный подход, выходной перечень, изменение усилия, тип слияния, различная параметризация, неубывающая зависимость, *endpoint graph*, исключение обмена, стабилизация установки, идея планирования, остов графа, подчиненный фрейм, график управляющего
Южный федеральный университет	метод стыковки, вставочный нейрон, движение дирижабля, трансформация фреймов, экспертное задание, виртуальная популяция, оптимизация, regulators synthesis, модульная система, синтез, мощность управляемость, синхронизация, барражирование, топология критериев, естественная декомпозиция, модель сао, функциональный-регулярный, маршрутизация коммивояжера, процедура преднастройки, сенсомоторная координация, неточный поиск, темпоральная теория, индуктивная программа, *impact*
Институт математики им. С.Л. Соболева СО РАН	обнаружение закона, семантический вывод, частотные триграммы, булевое программирование, *discovery* система, числовое представление, триграмма символов, идентификация авторства, результирующее отношение, природа стратегии, логический путь, открытие теории, машинный метод, адекватность существования, индикаторная функций, атомарное предсказание, эвристика, *regression trees*

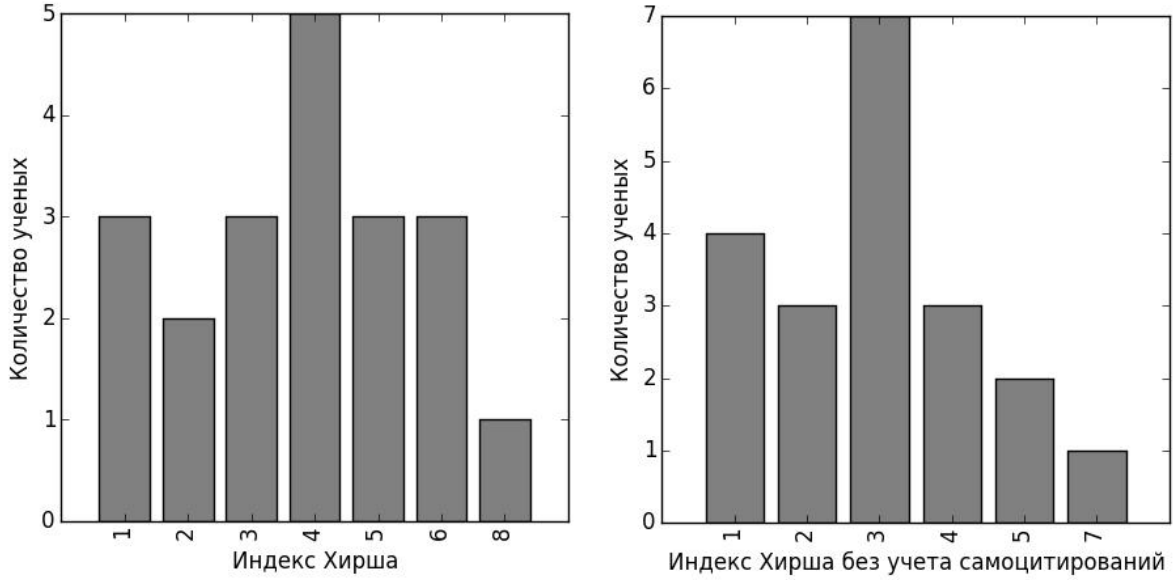


Рис. 7. Индекс Хирша ученых ФИЦ «Информатика и управление» РАН, работающих в области искусственного интеллекта

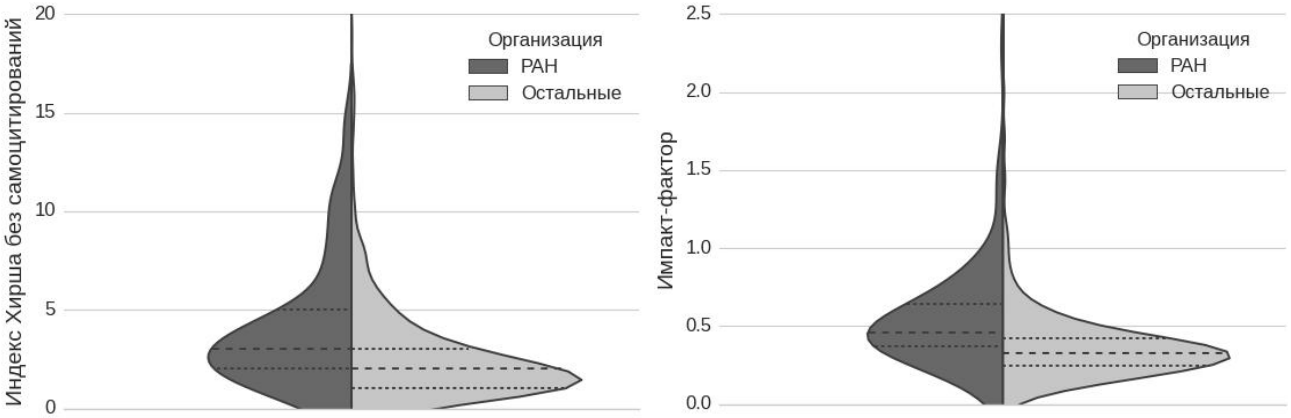


Рис. 8. Распределение индекса Хирша без учета самоцитирований (слева) и импакт-фактора (справа) для ученых, работающих в области искусственного интеллекта из институтов РАН и других организаций

СРАВНИТЕЛЬНЫЙ АНАЛИЗ ЦЕНТРОВ КОМПЕТЕНЦИИ В ОБЛАСТИ ИИ

Информация, полученная из профилей РИНЦ, использовалась для детального анализа центров компетенции. Например, из диаграммы (рис. 7) видно, что средний индекс Хирша ученых ФИЦ «Информатика и управление» РАН выше средних показателей в области искусственного интеллекта. Этот вывод статистически подтверждается U-тестом Манна-Уитни с критическим значением p=0,05 [8].

В рамках настоящего исследования организации из отфильтрованного рейтинга были также разбиты на 2 группы в зависимости от ведомственной принадлежности –институты РАН и другие организации (рис. 8).

Статистический U-тест Манна-Уитни показал, что в среднем индекс Хирша исследователей в области искусственного интеллекта из организаций РАН больше аналогичного показателя ученых из других организаций ($p=7*10^{-6}$). Импакт-фактор журналов, в которых публикуются сотрудники институтов РАН, в среднем также выше ($p = 2*10^{-10}$).

ЗАКЛЮЧЕНИЕ

Предлагаемая методика выявления центров компетенции позволяет устанавливать основные специальности, по которым защищаются диссертации в рассматриваемой предметной области, строить рейтинги организаций по наукометрическим показателям активных ученых, а также определять тематику исследований этих центров компетенции. Методика позволяет проводить анализ на уровне как полного списка организаций, так и подгрупп организаций, а также отдельных организаций, что может быть полезно при детальном исследовании центров компетенции.

Апробация выполнена на предметной области «искусственный интеллект», однако предложенная методика применима для анализа любой предметной

области, которая слабо представлена в западных базах цитирования и для которой в классификаторах отсутствуют соответствующие ей коды.

Основные направления дальнейших исследований – это применение представленной методики для анализа других предметных областей и создание алгоритмов выявления ведущих ученых в отдельных направлениях исследований. В ходе совершенствования представленной методики планируется применить тематические модели при построении ключевых слов для центров компетенции, что должно повысить качество выделения лексики для центров с широкой специализацией. Для повышения точности выявления центров компетенции будет использоваться дополнительная информация, предоставляемая зарубежными базами данных.

Еще одно перспективное направление работы – это применение методов анализа полных текстов для сопоставления научно-технических документов из разнородных источников (баз научных статей, описаний изобретений и др.), которые невозможно автоматически связать друг с другом из-за отсутствия сквозной классификации. Такое сопоставление было бы полезным при проведении комплексного анализа состояния дел в любой области науки и техники.

СПИСОК ЛИТЕРАТУРЫ

1. Еременко Г. Во всем виноват РИНЦ? // Троицкий вариант. – 2014. – № 163.– С. 7.
2. Фрадков А. РИНЦ продолжает врать // ТрВ-Наука. – URL: http://trv-science.ru/2015/09/08/risc-prodolzhaet-vrat/ (дата обращения: 15.05.2017)
3. Зибарева И.В., Солошенко Н.С. Тематическая структура российского сегмента научных журналов в глобальных и национальных информационных ресурсах // Третья международная конференция НЭИКОН «Электронные научные и образовательные ресурсы: Создание, продвижение и использование». Материалы конференции. – М.: НП НЭИКОН, 2015. – С. 255-259.
4. Осипов Г.С. Искусственный интеллект: состояние исследований и взгляд в будущее // Новости искусственного интеллекта – 2001. – Т. 43, №. 1.
5. Суворов Р.Е., Соченков И.В. Определение связанности научно-технических документов на основе характеристики тематической значимости // Искусственный интеллект и принятие решений. – М.: ИСА РАН, 2013.– №. 1. – С. 33-40.
6. Manning C.D. et al. Introduction to information retrieval. – Cambridge : Cambridge university press, 2008. – Vol. 1, №. 1. – P. 496.
7. Devyatkin D. et al. Full-Text Clustering Methods for Current Research Directions Detection // DAMDID/RCDL. – 2015. – P. 152-156.
8. Mann H.B., Whitney D.R. On a test of whether one of two random variables is stochastically larger than the other // The annals of mathematical statistics. – 1947.– P. 50-60.

Материал поступил в редакцию 29.08.17.

Сведения об авторах

ДЕВЯТКИН Дмитрий Алексеевич – младший научный сотрудник, Федеральный исследовательский центр «Информатика и управление» Российской академии наук, Москва
e-mail: devyatkin@isa.ru

СУВОРОВ Роман Евгеньевич – младший научный сотрудник, Федеральный исследовательский центр «Информатика и управление» РАН, Москва
e-mail: rsuvorov@isa.ru

ТИХОМИРОВ Илья Александрович – кандидат технических наук, доцент, зав. лабораторией, Федеральный исследовательский центр «Информатика и управление» РАН, Москва
e-mail: tih@isa.ru

УДК 025.43

А.Б. Антопольский, В.Н. Белоозеров, Н.Е. Калёнов, Н.Н. Шабурова, М.М. Якшин

Разработка семантической сети ключевых слов на основе дефинитивных связей*

Описано обоснование и методика создания сети связанных по смыслу терминов индексирования научных данных на основе анализа определений в системе терминологических словарей по всем областям науки и техники. Предусмотрена обработка определений в автоматическом и интеллектуальном режимах. Между терминами устанавливаются связи, соответствующие отношениям понятий в информационно-поисковых тезаурусах.

Ключевые слова: *терминологические словари, определения терминов, смысловые связи, тезаурусные отношения понятий, семантическая сеть, онтология информационного пространства*

За последние десятилетия основным направлением разработки логико-лингвистического обеспечения информационного поиска, технологий автоматической обработки текста и управления знаниями стали онтологии (например, [1]). Принципам их создания и использования посвящено большое количество публикаций. Только после 2000 г. появилось более 300 отечественных научных статей, связанных с теоретическими и практическими аспектами создания онтологий и их применения в информационных системах различных классов. Библиографические описания этих статей содержатся в указателе [2], где по этой проблеме выборочно представлена зарубежная литература, которая гораздо более обширна, чем отечественная.

Наиболее заметными отечественными обобщающими изданиями по созданию онтологий являются монографии Н.В. Лукашевич с соавторами [3] и В.Ш. Рубашкина [4]. В последней книге содержится подробный обзор как интеллектуальных, так и автоматизированных методов построения и пополнения онтологий. Кроме того, в специальной работе [5] В.Ш. Рубашкин и В.А. Капустин описали метод полуавтоматизированного формирования онтологий на основе энциклопедических и терминологических словарей.

В предыдущих наших публикациях [6, 7] было показано, что логико-лингвистической основой построения единого российского электронного пространства знаний должна быть онтология, построенная на лексике и парадигматике информационных языков, практически используемых в рамках этого пространства. Очевидно, что такими информационными языками, прежде всего, являются классификаторы научно-технической информации и библиотечные классификации. На их основе была создана терминологическая база данных *TERMIN*, реализованная в БЕН РАН на платформе SCIRUS [8]. Описание процедур создания этой базы данных содержится в работе [9].

Как показывает обзор методик построения онтологий, сделанный в приведенных выше работах, необходимым их компонентом является сеть терминов, связанных смысловыми отношениями. Эта семантическая сеть моделирует онтологию пространства знаний. При отображении пространства научного знания сеть отношений должна строиться на материале научных терминов, используемых для описания (индексирования) информационных ресурсов. Необходимо иметь в виду, что существующие традиционные классификационные системы описания (индексирования) научных информационных ресурсов концептуально и формально отстают от развития науки и информационной практики. Собственно, и сама таксономия наук в современной научной картине мира не совсем соответствует реальному положению дел. Появление новых научных отраслей, усиление междисциплинарных связей, развитие фундаментальной и прикладной науки, современные инфо- и инфраструктуры научных ресурсов (идеология открытой науки, связанных данных, больших данных, семантического веба и т. д.) требуют, на наш взгляд, разработки нового классификационного инструментария с учетом современных потребностей и технологий. Развитие науки, научной информации, интеграция в мировую среду сопровождаются развитием способов научных коммуникаций. Основой таких коммуникаций несомненно является система терминов, специфичных для той или иной отрасли, отражающая её онтологию. Первый и необходимый этап разработки современной лингвистической базы для навигации и информационного поиска в научно-информационных системах – это мониторинг, семантический анализ

* Публикация подготовлена в рамках проекта РФФИ (РГНФ) «Интерактивная система создания и поддержки онтологии научного знания на базе динамического комплекса терминологических словарей», грант № 17-03-12013

терминологических словарей и их динамическая корреляция с традиционными классификационными схемами. Необходима визуализация всех типов связей между терминами, понятиями, денотатами, наглядно представляющая их взаимосвязи, т.е. представляющая онтологию отрасли понятийно-терминологической картой. Хорошей основой для этого служит база данных *TERMIN*.

Существующие и разрабатываемые компьютерные и программные комплексы, равно как и информационные системы, для которых они предназначены, могут дать ожидаемые или близкие к ожидаемым результаты только при условии исследования, проектирования и практической реализации интеллектуальных систем и самого социума, в составе которого функционирование этих средств осмысленно и оправданно. Всякая информационная система формируется и действует внутри той или иной конкретной социальной и предметной области и предназначается для решения некоторого заданного круга задач. Среди средств информационных систем выделяются логические (методы и модели объектов), технические (вычислительные, измерительные и др. устройства), семиотические (языковые и неязыковые знаковые системы) и информационные (конкретные сведения об объектах, их свойствах, отношениях, поведении). Способы их взаимодействия включают организационные структуры и модели, по которым устанавливаются и воспроизводятся отношения, входящие в эти структуры.

Качество функционирования любой информационной системы во многом зависит от уровня адекватного описания соответствующего фрагмента предметной области, достигаемого посредством соответствия информационной модели реальному состоянию. Для описания предметной области используют естественные языки и искусственные формализованные языковые средства. Основной задачей является получение формального (не зависящего от СУБД) описания предметной области, которая подлежит моделированию. При информационном моделировании пространства документов, представляющих предметную область, приходится иметь дело с некоторыми сущностями, отсутствующими в текстах документов в явном формализованном виде, но несущими собственное реальное значение. В данном случае основной задачей становится выделение смысла информации в тексте, т.е. восстановление отдельных объектов и их взаимосвязей, которые либо описаны, либо упомянуты, либо подразумеваются неявно (data mining). Онтология должна стать базой данных, в которой будет храниться информация об объектах, процессах и явлениях, описанных в текстах, и эти данные должны сопровождаться информацией о свойствах, качествах и взаимосвязях описанных объектов. Должно быть предусмотрено, что один и тот же объект может описываться с использованием различных слов (терминов), либо даже не описываться, а подразумеваться косвенно, как интуитивно ассоциируемый с реально названными объектами.

Вслед за проблемой восстановления объектов (событий) возникают проблемы восстановления взаимосвязей, отношений, характеристик и т.д. Но если в первом случае задача решается известными средствами – составление словарей объектов (событий), словарей синонимов и т.д., то задача восстановления связей, отношений, характеристик, особенно описываемых неявно и/или разрозненно, усложнена тем, что она не решается без применения интеллектуальных технологий, базирующихся на проработанных моделях естественного языка, моделях построения текстов, моделях мышления.

На протяжении многих лет традиционной для задач информационного поиска формой смысловых отношений терминов является информационно-поисковый тезаурус [10, 11], в котором в качестве дескрипторов (терминов, связанных смысловыми отношениями) выступают выражения, используемые для описания содержания информационных ресурсов (индексирования документов), т. е. ключевые слова. Между дескрипторами обычно устанавливают три вида отношений:

- «ассоциация», отражающая факт существенного пересечения множеств ресурсов (документов), заиндексированных данными терминами;
- «иерархия», отражающая факт полного (или почти полного) вхождения множества ресурсов, заиндексированных одним из данных терминов, во множество ресурсов, заиндексированных другим термином;
- «эквивалентность», отражающая факт существенного совпадения множеств, стоящих за каждым из данных терминов.

Именно так понимаются связи в информационно-поисковом тезаурусе, используемом классически – для первичного индексирования документов, когда этот тезаурус является развитием классификационного инструментария, где каждый элемент – это класс документов (например, [12]).

Цель индексирования документов ключевыми словами – обеспечение максимально полного и точного их поиска по запросам пользователей. Эта цель была бы легко достижима, если бы при индексировании и при поиске использовались одни и те же лингвистические средства. Однако на практике между этими процессами возникают существенные «противоречия» – индексирование документов осуществляют авторы или специально обученные сотрудники библиотек и информационных учреждений (возможно, с использованием тезаурусов), поиск же информации осуществляют пользователи, зачастую не подозревающие о существовании тезаурусов, использованных для индексирования документов. Сгладить эти противоречия призваны онтологии предметных областей, связывающие дескрипторы тезаурусов со всей совокупностью терминов данной предметной области.

Поскольку мы предполагаем использовать связи терминов для поиска по независимо назначенным ключевым словам, то связи терминов должны отражать отношения понятий, представленных в системе знаний индексаторов и пользователей. Поэтому в создаваемой онтологии следует рассмотреть возможность установления двух типов отношений терминов – отношения между классификационными рубриками (на основе пересечений множеств документов) и от-

ношения между независимыми ключевыми словами (на основе связи интенсионалов понятий). Эти отношения позволят при информационном поиске находить данные, описанные различными выражениями естественного языка, данные о деталях и атрибутах искомого объекта, о его окружении (контексте), а также выходить на классификационные рубрики, систематизирующие данные по стандартным классам и обеспечивающие полноту поиска по этим классам.

Для того чтобы сеть терминов исполняла роль инструмента навигации и поиска в пространстве информационных ресурсов, в ней должны быть представлены именно ключевые слова, используемые в той или иной рубрике и привязанные к этой рубрике специальным указанием связи. В сети могут (точнее – должны) содержаться термины, используемые заказчиками информационного поиска, но привязка которых к классификационным рубрикам отсутствует или неизвестна. Эти термины также должны входить в сеть связей с ключевыми словами.

Такое описание набора терминов совпадает с характеристикой терминологических словарей в БД *TERMIN*. Поэтому целесообразно положить эти словари в основу построения сети семантически связанных терминов, моделирующей онтологию информационного пространства научной и технической информации.

Словари представлены в виде реляционной базы данных, включающей следующие «объекты»:

- словарь;
- термин;
- определение термина;
- источник;
- пользователь.

Объект «Словарь» имеет обязательные атрибуты – «название», «код ГРНТИ».

Объект «Термин», связанный с объектом «Словарь» отношением n:n (многие ко многим – один термин может относиться ко многим словарям, один словарь связан со многими терминами), имеет обязательный атрибут «название» и факультативный – «код рубрики».

Объект «Определение термина» имеет обязательный атрибут «текст», связь типа n:1 с объектом «Термин» (один термин может иметь несколько определений, но конкретное определение относится только к одному термину) и связь типа 1:n с объектом «Источник» (это определение относится только к одному конкретному источнику, но из одного источника можно выбрать много определений различных терминов).

Объект «Источник» (компетентный документ, из которого взято определение термина) имеет обязательный текстовый атрибут «описание» и факультативный – «дополнительная информация», он связан отношением 1:n с объектом «Определение термина»

Код ГРНТИ – это одна из рубрик верхнего уровня ГРНТИ. Он связан с терминами, представляющими собой (главным образом) ключевые слова, используемые для индексирования документов, принадлежащих к соответствующей рубрике в информационных ресурсах. Совокупность этих терминов с относящимися к ним определениями составляет терминологический словарь по данной отрасли знания.

Превратить эту структуру в сеть терминов, связанных по смыслу, можно на основе анализа определений (дефиниций), имеющихся у каждого термина. Ручной интеллектуальный анализ смысла и связей терминов – дорогостоящая процедура, требующая времени и квалифицированных кадров. Существенно облегчить эту процедуру позволяет автоматический анализ определений.

Очевидно, что если в определении термина А употреблён термин Б, то эти термины связаны друг с другом по смыслу. Характер этой связи при этом не определяется, но мы можем утверждать, что если эти термины являются ключевыми словами информационного массива, то в совокупности сведений, заиндексированных одним из этих ключевых слов, имеется информация, релевантная другому ключевому слову. Поэтому мы можем установить между терминами А и Б неспециализированную ассоциативную связь А×Б, которую полезно использовать при расширенном поиске информации по каждому из данных ключевых слов. Эти неспецифицированные связи полезно подвергнуть верификации и уточнению путём интеллектуального анализа, при котором специалист может квалифицировать связи семантическими категориями в зависимости от имеющихся задач обработки информации. Эта процедура реализует простейший критерий поиска семантических связей терминов из числа, описанных в [5]. Она не заменяет свободный интеллектуальный поиск смысловых отношений, но может дать быстрый результат по всему имеющемуся массиву терминов.

Предлагаемая система ассоциативных связей на множестве словарей ключевых слов послужит каркасом, на котором может быть построена полноценная онтология сферы НТИ, предложенная в работах [6, 7].

Таким образом, мы предлагаем начать построение онтологии с автоматического установления связей терминов по их дефинициям – с установления «дефинитивных связей». Алгоритм этой операции состоит в том, что для каждого термина нашего массива проверяется факт вхождения его в каждое определение по всем терминологическим словарям. В случае обнаружения термина А в определении термина Б между ними фиксируется в базе данных рефлексивное (не транзитивное) отношение А×Б (а также Б×А).

Соответствующим образом меняется структура базы данных *TERMIN* – добавляется таблица связей, содержащая идентификаторы терминов и ссылки на связи между ними.

Другой метод автоматического анализа определений может состоять в установлении меры близости текстов определений по совпадению терминов в их составе. Этот метод подлежит детальной разработке и требует интеллектуального вмешательства для определения порога, при котором между терминами с близкими определениями будет фиксироваться наличие ассоциации. Метод сходства определений предполагается реализовать на последующих этапах построения онтологии пространства НТИ.

Должны быть рассмотрены и опробованы также методы автоматического определения специфицированных связей терминов, описанные В.Ш. Рубашкиным и В.А. Капустиным [5].

Массив автоматически установленных ассоциаций в виде списка пар ассоциированных терминов поступает на интеллектуальную обработку, в ходе которой связи между терминами квалифицируются как та или иная тезаурусная связь с возможным подразделением связей на более специфичные виды и с возможным отказом подтвердить наличие связи. Смысл установленных связей между терминами А и Б таков:

А=Б – термины А и Б справедливы для индекирования одних и тех же информационных источников,

А>Б – массив источников, для которых справедливо индексирование термином А, включает массив источников, для которых справедливо индексирование термином Б,

А<Б – массив источников, для которых справедливо индексирование термином Б, включает массив источников, для которых справедливо индексирование термином А,

А×Б – массивы источников, для которых справедливо индексирование данным термином, существенно пересекаются.

В этой системе термины, совпадающие побуквенно, но относящиеся к разным предметным областям следует рассматривать как разные сущности, что отражается вхождением терминов в разные словари, отнесённые ГРНТИ к разным областям знания. Соответственно областям знания помеченными оказываются и смысловые связи, среди которых выделяются связи терминов одной области знания и связи между терминами разных областей знания.

В результате установления связей терминов мы получим структуру, характеризующуюся следующими объектами (сущностями):

• область знания (по верхнему уровню ГРНТИ);

• терминология, принятая в данной области знания (массив терминов – ключевых слов);

• определения каждого термина с указанием на источники этих определений;

• связи между терминами внутри данной области знания;

• связи между терминами различных областей знания.

Автоматически установленные связи дополняются интеллектуальным поиском связей и заимствованием связей и лексики из информационно-поисковых тезаурусов, в частности, из разработанных ранее тезаурусов по электронике [13, 14], педагогике [15], водному транспорту [16].

На последующих этапах предполагается совместить семантическую сеть ключевых слов с системой соответствия рубрик библиографических классификаторов [17], в результате чего мы получим исчерпывающую модель (онтологию) связей метаданных для тематического поиска и навигации в информационном пространстве научного знания.

СПИСОК ЛИТЕРАТУРЫ

1. Гаврилова Т.А. Использование онтологий в системах управления знаниями / Бизнес Инжиниринг Групп, 2003. – URL: http://bigc.spb.ru/ publications/bigspb/km/use_ontology_in_suz.php
2. Информационные языки в XXI веке. Библиографический указатель / сост. А.Б. Антопольский, Т.С. Маркарова. – URL: http://www.systemling.narod.ru/ informat/bibliografiya-ontologij.docx.
3. Добров Б.В., Иванов В.В., Лукашевич Н.В., Соловьев В.Д. Онтологии и тезаурусы: модели, инструменты, приложения. – М.: БИНОМ.ЛЗ, ИНТУИТ.ру, 2012. – 173 с.
4. Рубашкин В.Ш. Онтологическая семантика. – М.: Физматлит, 2013. – 348 с.
5. Рубашкин В.Ш., Капустин В.А. Использование определений терминов в энциклопедических словарях для автоматизированного пополнения онтологий // XI Всероссийская объединенная конференция «Интернет и современное общество». – СПб., 2008.
6. Антопольский А.Б., Белоозеров В.Н., Маркарова Т.С. О разработке онтологии на основе классификаторов научной информации и терминологических словарей // Информационные ресурсы России. – 2017. – № 5.
7. Антошкова О.А., Белоозеров В.Н., Дмитриева Е.Ю., Шапкин А.В. Разработка онтологии НТИ на основе библиографических классификаций // XXI научно-практический семинар «Информационное обеспечение науки: новые технологии» (Таруса, 3 – 7 июля 2017 г.). – М.: БЕН РАН, 2017.
8. Якшин М.М. Развитие платформы Scirus // Информационное обеспечение науки новые технологии : сб. науч. тр. / ред. Н.Е. Калёнов, В.А. Цветкова – М.: БЕН РАН, 2015. – С. 203-207.
9. Калёнов Н.Е., Белоозеров В.Н. Формирование терминологических словарей по лексике классификационных систем // Научно-техническая информация. Сер. 1. – 2015. – № 3. – С. 60-70.
10. ГОСТ 7.25–2001 Система стандартов по информации, библиотечному и издательскому делу. Тезаурус информационно-поисковый одноязычный. Правила разработки, структура, состав и форма представления. – М., 2001. – 16 с.
11. ISO 25964-1:2011 Information and documentation – Thesauri and interoperability with other vocabularies – Part 1: Thesauri for information retrieval. – Genève: ISO, 2011.
12. Михайлов А.И., Чёрный А.И., Гиляревский Р.С. Основы информатики. – Изд. 2-е перераб. и доп. – М.: Наука, 1968. – 370 с.
13. Белоозеров В.Н., Шабурова Н.Н. Развитие тезауруса классификационных систем по физике полупроводников // Новые технологии в информационно-библиотечном обеспечении научных исследований : сб. науч. тр. – Екатеринбург: ЦНБ УРО РАН, 2010. – С. 287-298.
14. Белоозеров В.Н., Шабурова Н.Н. Тезаурус тематических рубрик по физике полупроводни-

ков // Рукопись депонирована в ВИНИТИ 2013-12-24 № 379-В2013.
15. Маркарова Т.С. Педагогический тезаурус как терминосистема образовательной отрасли // Материалы 8-й международной конференции «Актуальные проблемы информационного обеспечения науки, аналитической и инновационной деятельности» (Москва, 28-30 ноября 2012 г.). – М.: ВИНИТИ, 2012. – С. 128-135
16. НТП ВИНИТИ РАН 85-2012 Список основных ключевых слов для координатного индексирования документов по проблемам водных перевозок, технической эксплуатации и ремонта флота : Нормативно-техническое предписание / С. М. Резер, В. Н. Белоозеров, Л. А. Рыжова, И. М. Кочнев. – М.: ВИНИТИ РАН, 2012. – 127 с.
17. Антошкова О.А., Белоозеров В.Н., Дмитриева Е.Ю. Разработка базовых соответствий между ГРНТИ и другими классификационными системами // Информационное обеспечение науки: новые технологии: сб. тр. / ред. Н.Е. Калёнов, В.А. Цветкова. – М.: БЕН РАН, 2015. – С. 137-146.

Материал поступил в редакцию 29.08.17.

Сведения об авторах

АНТОПОЛЬСКИЙ Александр Борисович – главный научный сотрудник ИНИОН РАН, доктор технических наук
e-mail: ale5695@yandex.ru

БЕЛООЗЕРОВ Виктор Николаевич – ведущий научный сотрудник, кандидат филологических наук, ВИНИТИ РАН
e-mail: nomoip@viniti.ru

КАЛЕНОВ Николай Евгеньевич – директор БЕН РАН, доктор технических наук,
e-mail: benran.ru

ШАБУРОВА Наталья Николаевна – заведующая научной библиотекой ИПФ СО РАН, кандидат педагогических наук
e-mail: shaburova@isp.nsc.ru

ЯКШИН Михаил Михайлович – научный сотрудник БЕН РАН.
e-mail: benran.ru

УДК 025.48

О.А. Антошкова, В.Н. Белоозеров, Е.Ю. Дмитриева, О.В. Смирнова, А.В. Шапкин, Н.Н. Шабурова

О методике построения онтологии научно-технической информации в виде сети библиографических классификаций*

Изложена методика построения сети классификаторов, образующей многоаспектное представление онтологии научно-технической информации, а также проиллюстрирована методика исследования эффективности автоматического определения семантических связей классификационных рубрик.

Ключевые слова: *библиографические классификации, онтология информационных ресурсов, научно-техническая информация, совместимость информационных систем, лингвистическое обеспечение, информационный поиск, сеть классификационных систем*

Указом Президента РФ «Об утверждении Основ государственной культурной политики» [1] и Федеральным законом «О библиотечном деле» [2] предусмотрено формирование единого российского электронного пространства знаний. Порядок его формирования был предложен в проекте положения о Национальной электронной библиотеке (НЭБ) [3]. В сфере научно-технической информации (НТИ) это предполагает создание возможности через информационную сеть получать научные сведения по любой тематике. Для этого необходимо не только, чтобы сеть содержала базы знаний по всем областям науки, но и чтобы она имела средства сопоставления запроса с тематикой информационных ресурсов, имеющихся в этих базах знаний. Проект положения о НЭБ прямо требует связывания электронных документов в пространстве знаний с системами классификации, поиска и извлечения информации.

Однако в настоящее время классификационные системы, которые служат для поиска и извлечения сведений из разных информационных ресурсов, представляют собой совокупность плохо связанных друг с другом тематических рубрикаторов. Даже в пределах одной сферы – естественных наук – одна часть источников информации систематизирована по международной Универсальной десятичной классификации (УДК), а другая – по отечественной Библиотечно-библиографической классификации (ББК). При этом наиболее авторитетные реферативно-библиографические сведения в базах данных издательских фирм *Thomson Reuters* [4] и *Elsevier* [5] описываются своими оригинальными рубрикаторами. Аналогичным образом поступают отечественные фонды поддержки науки РФФИ [6], РНФ [7] и др.

Исследование системы соответствий классификаторов по науке и технике рубрикам Государственного рубрикатора научно-технической информации (ГРНТИ) показало, что с большой степенью полноты и точности из этой системы можно вывести прямые смысловые связи между классами различных классификаторов, сопоставленных с ГРНТИ. Это открывает возможность при минимальных затратах трудовых ресурсов построить сеть применяемых в информационной практике классификационных систем. Такая сеть позволит осуществлять навигацию и поиск в пространстве информационных сетей, переходя от одной информационной системы к другой по смысловым связям используемых классификаторов, получать многоаспектное представление онтологии научно-технической информации. В предполагаемую сеть классификаторов, наряду с ГРНТИ, входят: Универсальная десятичная классификация, Библиотечно-библиографическая классификация, классификаторы мировых библиографических систем (*Scopus*, *Web of Science*), классификаторы российских фондов поддержки науки (РФФИ, РНФ) и др.

В сфере отечественных научных коммуникаций связь различных информационных ресурсов в некоторой степени осуществляет Государственный рубрикатор научно-технической информации (ГРНТИ), принятый во многих системах для индексирования документов наряду с УДК или ББК. В результате работы, проведённой ВИНИТИ РАН по заданию Минобрнауки [8] создана возможность связи через ГРНТИ с набором основных классификационных систем, важных для систематизации знаний и мониторинга научных исследований в России. Это классификаторы российских фондов поддержки науки,

* Публикация подготовлена в рамках проекта РФФИ «Исследование системы классификаторов по науке и технике, и разработка механизма смысловой навигации и поиска знаний в информационных сетях» – грант № 17-07-00153

мировых библиографических баз данных *Scopus* и *Web of Science* и др. В Системе классификационных схем ВИНИТИ РАН [9] установлена связь с ГРНТИ для классификаторов следующих систем: *Scopus, Web of Science (WoS),* Организации экономического сотрудничества и развития (ОЭСР), Всероссийской аттестационной комиссии (ВАК), Российского гуманитарного научного фонда (РГНФ, теперь подразделение РФФИ), Российского научного фонда (РНФ), Российского фонда фундаментальных исследований (РФФИ), Федерального агентства научных организаций (ФАНО), а также с большими классификациями знаний – Универсальной десятичной классификацией, Библиотечно-библиографической классификацией и Международной патентной классификацией.

Однако свободная навигация по информационным ресурсам требует наличия непосредственных связей между средствами входа в их базы данных, т. е. между их системами классификации. Такие связи можно получить в ряде случаев логическим выводом из соответствий рассматриваемых классификаций одинаковым рубрикам ГРНТИ. Поскольку такой вывод может быть сделан не во всех случаях, а тогда, когда он возможен, результирующая связь оказывается ослабленной и возникает вопрос о практической целесообразности использования связей с ГРНТИ для установления непосредственных связей классификационных систем. Для ответа на этот вопрос был проведён ряд экспериментов по сравнению результатов алгоритмического вывода о соответствии рубрик классификаторов через связь с ГРНТИ с результатами экспертной оценки такой связи. Полный просмотр больших классификационных систем (таких как УДК и ББК, содержащих до 100 тыс. позиций) в ручном режиме невозможен. Поэтому для эксперимента были взяты отдельные тематические фрагменты, выделенные по рубрикам ГРНТИ, приведённым в табл. 1.

Результаты этих экспериментов позволяют сделать следующие выводы:

- для больших классификаций (УДК и ББК) число полученных алгоритмически прямых связей определяется числом рубрик ГРНТИ данной тематики с уменьшением примерно на 10%. Следовательно, для полного объёма классификационных таблиц алгоритмически можно получить порядка 7 тыс. соответствий, включающих все классы верхнего уровня;
- для рубрикаций баз данных и фондов поддержки науки удаётся алгоритмически установить соответствия не менее чем для 80% рубрик;
- все алгоритмически установленные соответствия являются истинными. Не зафиксировано ни одного случая, когда алгоритмически установленная связь грубо противоречила бы интеллектуальному анализу. Случаи, когда имеется возможность интеллектуальным анализом уточнить алгоритмически установленную связь, составляют не более 20% всех связей;
- качество установленной сети связей можно оценить цифрами, указанными в табл. 2:

Данные табл. 2 свидетельствуют, что алгоритмическое установление прямых связей между классификациями на основе их соответствий ГРНТИ позволяет надеяться на получение системы перекрёстных связей классов различных систем «с точностью до ГРНТИ», что является приемлемым уровнем для соответствия поисковых массивов на данном этапе развития системы. Дальнейшее уточнение системы соответствий классификаторов возможно в ходе и на основе её практической эксплуатации путём интеллектуального анализа специалистами.

Таблица 1

Сводка экспериментов по эффективности алгоритмического вывода связи рубрик

Тематика	***Рубрики ГРНТИ***	***Сопоставляемые классификации***
Языкознание	16	УДК – ББК
Физика твёрдых тел	29.19	УДК – ББК
Физические основы электроники	29.35	УДК – ББК
Радиоэлектроника	47	УДК – ББК
Робототехника и интеллектуальные системы	55.30; 28.23; 27.47.23	WoS, Scopus, ОЭСР, РНФ, РФФИ друг с другом
Добыча и переработка нефти и газа	38.53; 52.47; 61.51	WoS, Scopus, ОЭСР, РНФ, РФФИ с УДК

Таблица 2

Ожидаемые показатели качества автоматического вывода сопоставлений классификационных систем

Показатель	***Значение показателя***
Объём по всем 10 сопоставляемым классификациям	до 30 тыс. и более связей
Полнота охвата рубрик связями	выше 80%
Точность совпадения с мнением экспертов	выше 75%
Точность истинности связи	выше 95%

Сеть связей между классификационными рубриками реализует тематическую структуру пространства информационных ресурсов и тем самым отражает онтологию (бытие) знаний в аспекте содержания документов, описанных в фондах той или иной из рассматриваемых классификаций. Заметим, что отношения классификационных рубрик не являются онтологией предметов изучения в научных трудах, они только косвенно связаны с сущностью решаемых научных задач, поскольку предназначены лишь для группировки источников знания по содержательной близости, не вдаваясь в природу этой близости. Мы будем рассматривать систему соответствия рубрик библиографических классификаций как представление **онтологии информационных ресурсов**. Построенная таким способом онтология является обобщением и развитием классификационного подхода к описанию содержания научных данных и позволяет рассматривать научную информацию под разными углами зрения, заложенными в разных классификационных системах.

Между рубриками различных классификационных систем мы устанавливаем те же связи, которые традиционно отражают библиографические классификации. В предлагаемых нами таблицах применяются следующие обозначения связи рубрик А и Б:

- А = Б рубрики совпадают по содержанию
- А < Б рубрика Б включает рубрику А
- А > Б рубрика А включает рубрику Б
- А × Б содержание рубрик пересекается
- А # Б отношение рубрик не установлено

Таким образом, онтология информационных ресурсов представляется как множество тематических рубрик, между которыми заданы отношения трёх видов: эквивалентности (транзитивное симметричное), иерархии (транзитивное антисимметричное) и пересечения тематики (симметричное). Эта структура удовлетворяет определению понятия «онтология», как оно было введено в информационную теорию Томом Грубером в 1991 г. [10]. Рассмотрение классификационных систем как частного случая онтологии можно найти, например, в книге [11]. В этой работе авторы противопоставляют онтологии (как описания предметной области) тезаурусам (как языковым структурам) внешним для предметной области (принадлежащим к сфере пользовательского интерфейса). В нашем случае между элементами онтологии (рубриками) установлены отношения, типичные для классических информационно-поисковых тезаурусов, но носителями этих отношений являются не элементы естественного языка (слова), а рубрики, в идеале представляющие определённые классы информационных ресурсов и, следовательно, принадлежащие к предметной области. Такой «тезаурус классификационных систем» был предложен нами ранее для частной тематики физики полупроводников [12]. В глобальной же сфере информационных ресурсов использование для тематического описания классификационных рубрик, уже представленных как атрибуты поисковых массивов, позволяет более прямым образом описать предметные области и не потерять при этом связь с языком пользователя, поскольку сами рубрики являются до некоторой степени определениями содержащихся в них сведений на естественном языке.

Исходя из вышеизложенного, мы предлагаем следующий план работ по построению онтологии ресурсов научно-технической информации.

Первый шаг. Выбор подлежащих сопоставлению классификационных систем. Сопоставлению несомненно подлежат две большие библиографические классификации, законодательно принятые для индексирования (описания тематики) всей издаваемой литературы – международная УДК и отечественная ББК. Для навигации по ресурсам отдельных видов документов могут понадобиться остальные классификации, автоматическая обработка которых не вызывает проблем. Дополнительный интеллектуальный анализ требуется только для ББК и МКИ, связи которых с ГРНТИ установлены не в полном объёме.

Второй шаг. Автоматический вывод прямых соответствий рубрик выбранных классификаций друг с другом на основе их соответствий одной и той же рубрике ГРНТИ.

Вывод производится по алгоритму пересечения отношений, описанному ранее в [9].

При наличии у двух рубрик К1 и К2 из двух разных классификаций связи с одной и той же рубрикой Г из ГРНТИ между рубриками К1 и К2 выводится непосредственная связь, определяемая по табл. 3, т. е. выбирается наименее строгая связь, совместимая с исходными отношениями к рубрике Г.

В отдельных случаях эта процедура приводит к установлению для рубрик К1 и К2 альтернативных вариантов связи. В этих случаях производится объединение этих альтернатив согласно табл. 4, т. е. выбирается та альтернатива, которая даёт более точное определение характера связи рубрик.

Таблица 3

Операция пересечения отношений рубрик

Результирующая связь К1 и К2		Связь К2 и Г:			
		=	<	>	×
Связь К1 и Г:	=	=	<	>	×
	<	<	<	#	#
	>	>	×	>	×
	×	×	×	#	#
Знаком # отмечены случаи, когда алгоритмически установить связи невозможно					

Таблица 4

Операция объединения отношений рубрик

Результирующая связь К1 и К2		Вторая альтернатива			
		=	<	>	×
Первая альтернатива	=	=	=	=	=
	<	=	<	=	<
	>	=	=	>	>
	×	=	<	>	×

Выявленные перекрёстные связи рубрик фиксируются в базе данных Системы классификационных схем ВИНИТИ [13, 14].

Третий шаг. Интеллектуальное редактирование связей. Алгоритмическая процедура не гарантирует выявление всех содержательных связей между рубриками разных классификаций, и алгоритмически выявленные связи не всегда точно описывают соотношение рубрик. Однако, как показывают проведённые эксперименты, потребность в уточнении полученной системы соответствий возникает буквально в единичных случаях. Поэтому мы считаем, что редактирование связей можно проводить главным образом в ходе эксплуатации системы на основе полученного опыта навигации по информационным ресурсам.

Четвёртый шаг. Наполнение классификационной онтологии лексикой ключевых слов. Ключевые слова являются альтернативным инструментом описания содержания документов. Мы имеем в виду те ключевые слова, которые уже приписаны документам информационных ресурсов. Каждое ключевое слово можно рассматривать и как определённую рубрику, содержащую все документы, заиндексированные данным ключевым словом. Это позволяет включать ключевые слова в общую сеть отношений, наряду с классификационными рубриками. Добавление этих ключевых слов позволит связать классификационные рубрики с естественным языком поисковых запросов, используемым потребителями информации. При этом мы остаёмся в пределах сферы онтологических признаков информационных ресурсов, поскольку ключевые слова взяты именно из реальных поисковых образов, являющихся встроенными метаданными в массиве документов.

Работа по выявлению ключевых слов и их определений частично проведена в рамках проекта Минобрнауки 2014-1016 гг. [8,15]. Работу по объединению данных по связям классификационных рубрик и ключевых слов предполагается начать после реализации системы прямых связей библиографических классификаций, проводимой ныне.

Иллюстрацией к методике определения эффективности алгоритмического установления связей между классификациями служат работы по сопоставлению УДК и ББК, выполненные по тематике исследований Института физики полупроводников СО РАН.

Ранее мы рассматривали фрагменты УДК и ББК, соответствующие разделу ГРНТИ ***47 Электроника. Радиотехника*** [13], и пришли к описанным здесь выводам, но эти выводы были сделаны на ограниченном материале.

Теперь мы дополняем анализ радиоэлектронной тематики тематикой физических основ полупроводниковой техники, т. е. разделом ГРНТИ ***29.19 Физика твёрдых тел***, содержащим 27 рубрик.

В УДК этим рубрикам в разной степени соответствует 139 классов (включая комбинированные). В ББК раздел естественных наук в настоящее время находится на стадии пересмотра, но недавно опубликована сокращённая версия таблиц ББК [14], которая содержит этот раздел в довольно подробном виде. Мы будем устанавливать соответствия ГРНТИ именно с этим вариантом ББК. Надеемся, что, когда пересмотренные таблицы по естественным наукам будут полностью опубликованы, внесение корректив в таблицу соответствия классификаций не составит большого труда.

Всего найдено 14 классов ББК, соответствующих рубрикам ГРНТИ. Сводная таблица содержит 174 тройственных соответствия УДК – ГРНТИ – ББК, её фрагмент приведён в *Приложении 1*. В полном виде она представлена на сайте: http://systemling.narod.ru/UDC-GRNTI-BBC/29-19-source.docx.

Из этой тройственной таблицы следует исключить среднюю часть – ГРНТИ, выводя прямые отношения УДК – ББК по табл. 3 и 4, и мы получим таблицу прямых отношений, фрагмент которой приведён в *Приложении 2*, а в полном объёме она представлена в свободном доступе на сайте: http://systemling.narod.ru/UDC-GRNTI-BBC/29-19.docx.

Таким образом, в ходе наших работ выявлено и верифицировано около 650 соответствий классов УДК и ББК по основной тематике Института физики полупроводников СО РАН – вопросам физики твёрдого тела и электроники. При этом подтверждён вывод о целесообразности применения автоматической процедуры установления прямых отношений классификационных систем.

ISSN 0548-0019. НТИ СЕР. 1. ОРГ. И МЕТОДИКА ИНФОРМ. РАБОТЫ. 2017. № 11

ПРИЛОЖЕНИЕ 1

Таблица сопоставления классификаций УДК – ГРНТИ – ББК (фрагмент)
по тематике раздела ГРНТИ *29.19 Физика твёрдых тел*

УДК			ГРНТИ		ББК. Сокращённые таблицы			
Код	Наименование	Связь	Код	Наименование	Связь	Коды		Наименование
538.9	**Физика конденсированного состояния (жидкое и твёрдое состояние)**	X	**29.19**	**Физика твёрдых тел**	=	В37	22.37	Физика твёрдого тела. Кристаллография
539	**Строение материи**	>	**29.19**	**Физика твёрдых тел**	=	В37	22.37	Физика твёрдого тела. Кристаллография
539.2	Свойства и структура молекулярных систем	X	**29.19**	**Физика твёрдых тел**	=	В37	22.37	Физика твёрдого тела. Кристаллография
539.3	Механика деформируемых тел. Упругость. Деформации	<	**29.19**	**Физика твёрдых тел**	=	В37	22.37	Физика твёрдого тела. Кристаллография
539.4	Прочность. Сопротивляемость	<	**29.19**	**Физика твёрдых тел**	=	В37	22.37	Физика твёрдого тела. Кристаллография
548	**Кристаллография**	<	**29.19**	**Физика твёрдых тел**	=	В37	22.37	Физика твёрдого тела. Кристаллография
548	**Кристаллография**	X	29.19.01	Общие вопросы [физики твёрдых тел]	<	В37	22.37	Физика твёрдого тела. Кристаллография
.....								
537.622.2	Диамагнитные материалы	=	29.19.47	Диамагнетики	<	В334	22.334	Магнетизм
537.622.2	Диамагнитные материалы	=	29.19.47	Диамагнетики	X	Ж37	30.37	Материалы с особыми свойствами
537.611.4	Теория отдельных видов магнетизма	X	29.19.49	Ядерный магнетизм	<	В37	22.37	Физика твёрдого тела. Кристаллография
537.611.4	Теория отдельных видов магнетизма	X	29.19.49	Ядерный магнетизм	<	В334	22.334	Магнетизм
537.611.4	Теория отдельных видов магнетизма	X	29.19.49	Ядерный магнетизм	X	В383	22.383	Физика атомного ядра
539.143.43	Магнитный момент …	X	29.19.49	Ядерный магнетизм	X	В383	22.383	Физика атомного ядра
539.143.43	Магнитный момент …	X	29.19.49	Ядерный магнетизм	<	В334	22.334	Магнетизм

ПРИЛОЖЕНИЕ 2

Таблица прямых сопоставлений УДК и ББК (фрагмент)
по тематике раздела ГРНТИ *29.19 Физика твёрдых тел*

УДК		Связь	ББК. Сокращённые таблицы		
Код	Содержание	В скобках – решение эксперта по уточнению связи	Коды		Содержание
53.08	Общие основы и теория измерений. …	X	В37	22.37	Физика твёрдого тела. Кристаллография
531.7	… Методы и единицы измерений	X	В37	22.37	Физика твёрдого тела. Кристаллография
		X	В37в6	22.37в6	Физика твёрдого тела. Кристаллография - Методы
536.42	Влияние подвода или отвода тепла на объём и структуру тел. Фазовые переходы	X	В37	22.37	Физика твёрдого тела. Кристаллография
537.222.2	Электрический заряд в полупроводниках	<	В379	22.379	Физика полупроводников и диэлектриков
		X (<)	Ж37	30.37	Материалы с особыми свойствами
537.226	Электрические свойства диэлектриков	<	В379	22.379	Физика полупроводников и диэлектриков
537.226.4	Сегнетоэлектричество	X	В33	22.33	Электричество и магнетизм
		X (<)	В37	22.37	Физика твёрдого тела …
		X	Ж37	30.37	Материалы с особыми свойствами
537.226.4 :539.21	Свойства твёрдого тела – Сегнетоэлектричество	<	В37	22.37	Физика твёрдого тела. Кристаллография
537.311.32	Сопротивление и проводимость в диэлектриках	<	В379	22.379	Физика полупроводников и диэлектриков
537.311.322	Электрический ток в полупро-водниковых материалах	<	В379	22.379	Физика полупроводников и диэлектриков
		X (<)	Ж37	30.37	Материалы с особыми свойствами
537.61	Теория магнетизма	X	В37	22.37	Физика твёрдого тела. Кристаллография
		X (<)	В334	22.334	Магнетизм
537.61:539. 21	Свойства твёрдого тела – Теория магнетизма	<	В37	22.37	Физика твёрдого тела. Кристаллография
		<	В334	22.334	Магнетизм
537.611.4	Теория отдельных видов магнетизма	X	В37	22.37	Физика твёрдого тела. Кристаллография
		X	В334	22.334	Магнетизм

СПИСОК ЛИТЕРАТУРЫ

1. Указ Президента Российской Федерации от 24.12.2014 № 808 «Об утверждении Основ государственной культурной политики». – URL: http://www.kremlin.ru/acts/bank/39208. http://www.kremlin.ru/acts/bank/39208
2. Федеральный закон Российской Федерации от 29.12.1994 № 78-ФЗ «О библиотечном деле» (в редакции от 03.06.2016). – URL: http://fzrf.su/zakon/o-bibliotechnom-dele-78-fz/st-18.1.php.
3. Положение о Национальной электронной библиотеке: 2 вариант. – URL: www.unkniga.ru/images/docs/ 2017/polozhenie-neb-2-variant.pdf.
4. Clarivate Analytics. Science Citation Index Expanded. Scope Notes. 2017. – URL: http://ip-science.thomsonreuters.com/mjl/scope/scope_scie/#AA.
5. Elsevier. All Products. Books & Journals. All Subject Areas. – URL: https://www.elsevier.com/catalog.
6. Классификатор РФФИ для конкурсов 2017 года // Российский фонд фундаментальных исследований. Конкурсная документация. – URL: http://www.rfbr.ru/rffi/ru/contest_documents.
7. Российский научный фонд. Классификатор Фонда. – URL: http://rscf.ru/ru/classification.
8. Сопоставление ГРНТИ с другими классификационными системами с целью совершенствования системы тематической кодификации НИР, НИОКР гражданского назначения. Формирование системы соответствий между различными классификаторами в сфере научно-технической информации : Заключительный отчёт по соглашению ВИНИТИ РАН и Минобрнауки России № 14.601.21.0001 (шифр проекта 2014-14-573-0024-001). – М.: ВИНИТИ, 2015.
9. Шапкин А.В. Практические вопросы построения системы классификационных схем // Научно-техническая информация. Сер. 2. – 2006. – № 6. – С. 1-13.
10. Gruber T.R. The role of common ontology in achieving sharable, reusable knowledge bases // Principles of Knowledge Representation and Reasoning. Proceedings of the Second International Conference / eds. J.A. Allen, R. Fikes, E. Sandewell. – Morgan Kaufmann, 1991. – P. 601-602.
11. Добров Б.В., Иванов В.В., Лукашевич Н.В., Соловьёв В.Д. Онтологии и тезаурусы: модели, инструменты, приложения: учеб. пособ.. – М., 2012. – 173 с. – (Основы информационных технологий).
12. Белоозеров В.Н., Шабурова Н.Н. Сопоставительный тезаурус классификационных систем по физике полупроводников // Информационное обеспечение науки: новые технологии : сб. науч. тр. / ред. Н. Е. Калёнов. – М.: Науч. мир, 2009. – С. 311–322.
13. Белоозеров В.Н., Шабурова Н.Н. Метод сопоставления классификаций на основе соответствия рубрикам ГРНТИ (на примере УДК и ББК) // Научно-техническая информация. Сер. 2. – 2016. – № 10. – С. 13-24.
14. Библиотечно-библиографическая классификация. Сокращённые таблицы : практическое пособие / глав. ред. Э.Р. Сукиасян, РГБ, РНБ, БРАН. – М.: Пашков дом, 2015. – 672 с.
15. Белоозеров В.Н. Технология разработки терминологических словарей по лексике классификационных систем // Информационное обеспечение науки: новые технологии: сб. науч. тр. / ред. Н.Е. Калёнов, В.А. Цветкова. – М.: БЕН РАН, 2015. – С. 126-136.

Материал поступил в редакцию 29.08.17.

Сведения об авторах

АНТОШКОВА Ольга Александровна – заместитель заведующего отделением ВИНИТИ РАН, Москва
e-mail: oant@viniti.ru

БЕЛООЗЕРОВ Виктор Николаевич – кандидат филологических наук, ведущий научный сотрудник ВИНИТИ РАН
e-mail: nomoip@viniti.ru

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

СМИРНОВА Ольга Викторовна – научный сотрудник ВИНИТИ РАН
e-mail: typo@viniti.ru

ШАПКИН Александр Владимирович – кандидат технических наук, начальник Управления ВИНИТИ РАН
e-mail: ss@viniti.ru

ШАБУРОВА Наталья Николаевна – кандидат педагогических наук, зав. научной библиотекой Института физики полупроводников СО РАН, г. Новосибирск
e-mail: shaburova@isp.nsc.ru

УДК [025.4 : 81’371] : 005.96

А.Э. Анисимова, А.А. Рязанова, А.Ю. Щербаков

Семантическое ядро как универсальный инструмент классификации и систематизации неструктурированной информации в области человеческого капитала*

Рассматриваются методика и программный комплекс выделения семантического ядра текстового массива для обеспечения высокой степени точности отбора резюме на соответствующие вакансии, требования к которым заданы в произвольной текстовой форме. Такой отбор позволяет составить стабильную базу данных вакансий и соответствующую ей базу данных резюме. Наличие этих двух баз данных открывает путь для дальнейшего высокоэффективного автоматизированного анализа ключевых навыков, реализуемого на базе семантического ядра.

Ключевые слова: *человеческий капитал, большие данные, сайты по трудоустройству, информационный поиск, семантическое ядро, кластерный анализ, совместное употребление слов*

ВВЕДЕНИЕ

С точки зрения управления человеческим ресурсом основные усилия в процессе регулирования рынка труда на современном уровне направлены на устранение препятствий при поиске рабочего места и соответствующего работника. Можно назвать две основные причины, по которым квалифицированные специалисты не находят подходящего рабочего места. Первая из них связана с тем, что работник и работодатель описывают один и тот же круг навыков и компетенций с помощью различных терминов, используют разные наименования вакансий. Возможна, и даже более распространена, и обратная ситуация, когда одним термином (например, технолог) описываются принципиально разные специальности (например, технолог пищевого производства, технолог швейного оборудования, технолог инженерных коммуникаций и т.д.). В этом случае необходимо применять более сложные системы поиска, чем поиск по заголовкам объявлений, заложенный практически во все современные сайты по трудоустройству.

Одна из целей настоящего исследования — определение круга специальностей, представляющих наибольший интерес для абитуриентов с точки зрения будущего трудоустройства и заработной платы. Для составления списка основных специальностей важны автоматические инструменты, устанавливающие соотношение вакансий и резюме. Из списка резюме, имеющих подходящее название, на основе семантического анализа текста, предлагается удалять резюме, фактически не соответствующие требованиям работодателей. При превышении количества вакансий по сравнению с количеством резюме можно говорить о перспективности соответствующей специальности на рынке труда. При обратной ситуации делается вывод об угасании значения профессии или о несоответствии квалификации имеющихся специалистов требованиям, предъявляемым работодателем.

Технологические аспекты исследований в области развития человеческого капитала связаны со способами взаимодействия рынка труда, системы образования и технологической инфраструктуры, позволяющей находить специалистов, подходящих для определенной работы, а также обнаруживать системные лакуны в подготовке работников. Алгоритм взаимодействия в первом приближении можно представить в виде «черного ящика», где на входе имеются требования работодателя к кандидатам на вакансии, характеристики реальных работников и их резюме, а на выходе – вывод о соответствии квалификации работников требованиям рынка труда и, как следствие, – предложения для государственных или муниципальных органов профессионального образования в области повышения квалификации или переподготовки работников [1].

Автоматизированная обработка данных и поиск релевантной информации существенно упрощаются благодаря единым словарям терминов. Однако единый актуализированный словарь специальностей,

* Публикация подготовлена в рамках гранта, поддержанного РФФИ, грант № 16-33-01023 (введение, обзор литературы, применение комплекса индексирования и анализа текстов в исследовании сайтов по трудоустройству, изучение роли семантического ядра для выделения ключевых навыков) и в рамках работ, поддержанных РФФИ, грант № 15007-08522 (создание комплекса индексирования и анализа текстов, описание принципа работы программы).

узко характеризующих определенный род деятельности, еще только предстоит создать. Для подготовки блока объявлений из базы данных резюме, наиболее полно соответствующих объявлениям в базе данных вакансий, применима методика множественного семантического анализа текстов. При этом принципиальное значение имеет автоматизация процедуры анализа, предваряемая тестовыми исследованиями, необходимыми для определения уровня предварительной ручной обработки файлов.

ОБЗОР ЛИТЕРАТУРЫ

В научной литературе значительное место отводится публикациям, посвященным тематическому моделированию, с помощью которого возможно в значительных по объему текстах определять основные, характерные тематические блоки (topic modeling) [2]. Тематическое моделирование широко применяется при информационном поиске с целью обнаружения латентной нерубрицируемой информации, в ряде случаев – принципиально новой информации. Технология тематического моделирования позволяет свести текст, состоящий из миллионов терминов, к нескольким сотням тем на основе учета вероятности, с которой в той или иной тематике генерируются различные слова.

Наиболее распространенными методами тематического моделирования считаются латентное размещение Дирихле (LDA – latent Dirichlet allocation) и вероятностный латентно-семантический анализ (pLSI – probabilistic latent semantic indexing). Основы LDA были заложены Д. Блеем, Э. ЫНом и М. Джорданом в 2003 г. [3]. С точки зрения латентного размещения Дирихле служебные слова обладают равной вероятностью в различных тематиках, а любая тематика обладает вероятностью генерировать специфические слова. При наличии соответствующего рубрикатора, приписывающего конкретные слова определенной тематике, определить в любом тексте набор тем – достаточно простая задача. Этот метод текстуального анализа применяется в различных наукометрических исследованиях [4].

Приведенные методики чрезвычайно удобны для определения качественного сходства между двумя документами и выделения ключевых терминов, характеристически описывающих определенный текст. При наличии ключевых терминов путем анализа их попарного распределения в каждом из сегментов текста возможно предпринять кластерный анализ, результатом которого будет построение иерархической картины терминов. В рамках библиотечных и информационных наук метрические исследования стали крайне популярны в последние годы и включают анализ совместного употребления слов [5-7].

Все более широкое применение получают семантические системы на основе использования готовых словарей в сфере управления высшим образованием. Так сотрудники школы информационных систем Сингапурского университета менеджмента и специалисты в области преподавания информатики, провели совместное исследование, посвященное автоматической обработке учебных программ по информатике на основе анализа разработанного словаря профессиональных компетенций [8]. С определенного момента проблема полноты и целостности вузовского образования стала рассматриваться с точки зрения овладения компетенциями (competencies), необходимыми для начала трудовой деятельности. Компетентностный подход и использование в дизайне учебного курса понятия «результат обучения» привносят необходимую ясность и прозрачность в разрабатываемые обучающие курсы. С целью выявления необходимых компетенций возрастает роль систем семантической обработки больших текстов. Сущность анализа учебного плана заключается в оценке компонентов учебного плана и разработке предложений по его улучшению.

В рамках единого исследования в отдельных вузах США и Сингапура в 2008 г. специально разрабатывался список навыков в области информационных систем для бизнеса [9], который формировался путем сведения экспертных оценок с учетом мнения работодателей и личного опыта работы ИТ-специалистов. Для европейских вузов с 2014 г. уже существует единая квалификационная матрица по специальности «информатика», но в других странах общая классификация только еще формируется ручным способом путем удаления повторов и объединения близких навыков. Существуют и другие исследования, посвященные анализу учебных программ на основе совместного употребления терминов-компетенций [10].

С точки зрения развития национального человеческого капитала большое значение имеет аналитическая обработка быстроменяющихся данных на национальных сайтах по трудоустройству. В качестве частного примера можно привести исследование [11], посвященное анализу позитивного опыта соискательства на сайтах по трудоустройству Тайваня с помощью технологии решения прикладных проблем (task-technology fit - TTF) с использованием специализированного программного обеспечения (UTAUT2). Модель TTF получила распространение после публикации Д. Гудхью и Р. Томпсона в 1995, в которой обосновывалось преимущество технологий, создаваемых специально под конкретную задачу [12]. Существуют научно-популярные видео, поясняющие мощь и значение нового вида технологии (https://youtu.be/R9UGr5SpzlQ).

Аналогичная технология, ориентированная на решение конкретной задачи, – поиск актуальных навыков растущих профессий, – необходима для любой региональной сферы трудоустройства.

МЕТОДИКА ФОРМИРОВАНИЯ БАЗ ДАННЫХ НА ОСНОВЕ СЕМАНТИЧЕСКОГО АНАЛИЗА

Поскольку требования работодателя к кандидатам на вакансии, характеристики реальных работников и их резюме являются неструктурированными текстовыми данными, то содержание «черного ящика» может обрабатываться с помощью разного рода семантических алгоритмов. К ним относятся индексирование текстов, их сравнение, выявление смысла, статистический анализ текстов, поиск в текстах и др. В первом приближении задача поиска максимально пригодного

с точки зрения работодателя резюме сводится к сравнению двух неструктурированных текстов.

Задача сравнения двух текстов в информатике относится к классическим. К настоящему времени уже разработано достаточное количество методов семантического текстуального поиска, основанного на анализе частоты встречаемости слов, их соответствия определенной тематике, их совместного употребления и вероятностного распределения [13].

Если выбрать первое слово первого текста и сравнить его со всеми словами второго текста, то при нахождении можно констатировать, что это слово встречается и в первом, и во втором тексте, при ненахождении – что это слово есть только в первом тексте. После завершения процедуры во втором тексте остаются слова, которые встречаются только в нем. Трудоемкость этой процедуры составляет в среднем произведение длин текстов в словах на среднюю длину текста.

Для оптимального решения этой задачи используется аппарат небиективных (неоднозначных) отображений со следующими свойствами.

Пусть дано слово W произвольной длины L в некотором алфавите A. Рассмотрим преобразование H(W)=h, которое отображает слова произвольной длины в слово фиксированной длины.

Это преобразование должно обладать следующим свойством: при случайном равновероятном выборе двух слов W_1 и W_2 в алфавите A из множества возможных соответствующие им слова h_1 и h_2 должны быть с высокой вероятностью различны.

Если преобразование H является размешивающим преобразованием по Шеннону, то для оценки вероятности используют, как правило, длину слова h.

Предположим, что длина хеш-слова равна 3-м байтам. Тогда условная вероятность $P(h_1=h_2/W_1$ не равно $W_2)$ оценивается величиной порядка 2^{-24} т.е. 10^{-7} (с учетом того что 2^{10} =1024 – это приблизительно $1000=10^3$).

Следовательно, к произвольному тексту на любом языке можно применить следующее преобразование: каждое отдельное слово текста W, длиной более чем длина h, заменить значением (хеш-значением) функции H (хеш-функция) от него. Для общности можно заменять хеш-значением все слова, независимо от их длины.

В результате текст преобразуется в последовательность двоичных чисел, назовем их хеш-слова, каждое из которых будет длиной |h|, т.е. длиной хеш-значения (в приводимом примере 3 байта). Принципиальным результатом такого преобразования является то, что любые конструкции для сравнения и поиска становятся равной длины и нет необходимости сравнивать слова различной длины.

Далее, для каждого текста T_i одновременно с преобразованием его к хеш-словам строится словарь D_i, состоящий из неповторяющихся хеш-значений и соответствующих им слов.

Словарь D_i является мерой содержания текста, он позволяет не только оптимизировать поиск в тексте (ищем первоначально слова поискового запроса в словаре, при их наличии – ищем их в тексте, что и делают современные поисковые машины), но и сравнивать тексты, содержащие неструктурированные данные, между собой.

На основе описанного алгоритма работает используемый в настоящем исследовании комплекс индексирования и анализа текстов проф. А.Ю. Щербакова [14]. Программа индексирования текстов m_ind при запуске в формате m_ind[.exe] filename.ext создает три файла:

filename.csv – список слов (в кодировке Windows), встречающихся в индексированном тексте (словарь). Файл состоит из записей длиной 35 байт, из которых 32 байта занимают слова, дополненные пробелами, символ «;» и два символа перевода строки;

filename.lmd – файл индексов;

filename.num – файл двухбайтных значений, i-е поле равно количеству слов с номером i-й записи в словаре, встретившихся в индексируемом тексте.

Программа сравнения текстов tcmp при запуске в формате Tcmp[.exe] filename1.ext1 filename2.ext2 производит теоретико-множественное сравнение заданных в аргументах двух текстов (файлы filename1.ext1 и filename2.ext2 должны быть предварительно проиндексированы программой m_ind) и создает три файла:

- onlyone.csv – слова, встречающиеся только в первом тексте (filename1),
- onlytwo.csv – слова, встречающиеся только во втором тексте (filename2),
- common.csv – слова, встречающиеся как в filename1 так и в filename2.

Программа измеряет меру сходства текстов в файлах filename1.ext1 и filename2.ext2. Для этого определяются три меры [12].

Пусть T1 – число слов в первом тексте (filename1.ext1), T2 – число слов во втором тексте (filename2.ext2), O1 – число слов в файле onlyone.csv, O2 – число слов в файле onlytwo.csv, C – число слов в файле common.csv.

Тогда C – мощность (число элементов) пересечения двух множеств filename1.ext1 и filename2.ext2. Мера сходства текстов должна быть равна:

1) 0, если C=0, т.е. если пересечение множеств двух текстов пусто;

2) 1, если O1=O2=0 и тексты совпадают.

При этом выполняется равенство T1+T2=O1+O2+ +2C, это следует из равенств T1=O1+C и T2=O2+C.

Тогда можно определить меру R1=1/2(C/T1+ +C/T2). Как показывают эксперименты, эта мера будет принимать максимальные значения при оценке сходства текстов.

Кроме того, можно определить «естественную» меру R2=2C/(T1+T2), которая, как легко видеть, принимает значение 0, если тексты не совпадают (C=0), и 1, если C=T1=T2 в случае совпадения текстов.

Наконец, можно определить также третью, минимальную меру сходства текстов R3=C/(O1+O2+C).

Мера R3 необходима для тех случаев, когда при сравнении текстов из рассмотрения удаляется часть слов, например, слова малой длины (2 символа – предлоги, союзы, междометия) либо иные слова по выбору аналитика.

Проиллюстрируем изложенное конкретными примерами. Для этого возьмем одно типовое требование к вакансии и 4 резюме соискателей с сайта

hh.ru. Для анализа программой tcmp у нас имеется пять файлов: base_t, который содержит в неструктурированном произвольном виде требования работодателя к вакансии «инженер-технолог» по направлению «нефтедобыча и нефтепереработка»; резюме «Инженер-технолог» по направлению «нефтедобыча и нефтепереработка» первого соискателя – файл rt1; резюме «Инженер, Технолог» по направлению «нефтедобыча и нефтепереработка» второго соискателя – файл rt2; а также для иллюстрации технологии файлы двух резюме «Инженер-технолог швейного производства» – файлы rt3 и rt4. Приведем результаты попарного сравнения всех резюме с файлом base_t.

Для первого соискателя:

```
Min word length in COMMON => 0
Read pages.....
Success comparing! See onlyone,onlytwo and COMMON files
Files:
[rt1.txt]=398 words [base_t.txt]=171 words All=569
[onlyone]=347 [onlytwo]=120 [common]=51 All=569
Files metrics is correct
1-st Equal metric = 0.213193 [21%] ->Hihg
Null-Equal metric = 0.179262 [17%] ->Medium
2-d  Equal metric = 0.098456 [9%] ->Down
Medium = 0.163473 [16%]
```

Для второго соискателя:

```
Min word length in COMMON => 0
Read pages.....
Success comparing! See onlyone,onlytwo and COMMON files
Files:
[rt2.txt]=181 words [base_t.txt]=171 words All=352
[onlyone]=158 [onlytwo]=148 [common]=23 All=352
Files metrics is correct
1-st Equal metric = 0.130787 [13%] ->Hihg
Null-Equal metric = 0.130682 [13%] ->Medium
2-d  Equal metric = 0.069909 [6%] ->Down
Medium = 0.110349 [11%]
```

Для третьего соискателя:

```
Min word length in COMMON => 0
Read pages.....
Success comparing! See onlyone,onlytwo and COMMON files
Files:
[rt3.txt]=277 words [base_t.txt]=171 words All=448
[onlyone]=252 [onlytwo]=146 [common]=25 All=448
Files metrics is correct
1-st Equal metric = 0.118226 [11%] ->Hihg
Null-Equal metric = 0.111607 [11%] ->Medium
2-d  Equal metric = 0.059102 [5%] ->Down
Medium = 0.096215 [9%]
```

Для четвертого соискателя:

```
Min word length in COMMON => 0
Read pages.....
Success comparing! See onlyone,onlytwo and COMMON files
Files:
[rt4.txt]=315 words [base_t.txt]=171 words All=486
[onlyone]=275 [onlytwo]=131 [common]=40 All=486
Files metrics is correct
1-st Equal metric = 0.180451 [18%] ->Hihg
Null-Equal metric = 0.164609 [16%] ->Medium
2-d  Equal metric = 0.089686 [8%] ->Down
Medium = 0.144771 [14%]
```

В приведенных результатах работы программы по тексту «1-st Equal metric» – мера R1, «Null-Equal metric» – мера R2, «2-d Equal metric» – мера R3.

Как легко видеть, резюме первого соискателя максимально соответствует требованиям работодателя, первая мера принимает значение 21% (1-st Equal metric = 0,213193 (21%)). Для остальных соискателей первая мера принимает значения 13, 11 и 18%, соответственно.

Не составляет сложности объяснить тот факт, что резюме второго соискателя гораздо меньше соответствует требованиям вакансии: его опыт работы по направлению «нефтедобыча и нефтепереработка» составляет менее трех лет, в то время как опыт работы первого соискателя – более 18 лет, резюме гораздо меньше по объему и содержит меньшее количество навыков по специальности.

Неожиданно высокую меру сходства выявило сравнение четвертого резюме «Инженер-технолог швейного производства» с базовой вакансией (длина текста резюме сопоставима с длиной текста первого резюме), что свидетельствует о наличии большого количества слов, общих для большинства специальностей «инженер-технолог» в разных областях деятельности, таких как «производственные», «технологические», «продукция», «разработка», «подготовка», «составление» и др.

Для усовершенствования предлагаемых методов выполним следующее. Рассмотрим специальность «инженер-конструктор радиоэлектронной аппаратуры». Объединим все требования к вакансии «инженер-конструктор радиоэлектронной аппаратуры» в один текстовый файл и применим к нему преобразование m_ind. В результате получим словарь вакансии, содержащий свыше 500 позиций – базовых слов, описывающих требования к соискателю, хотя бы один раз встречающихся в объединенном тексте. Далее проведем ручную экспертную обработку текста – исключим предлоги и союзы, вспомогательные и малозначимые для рассматриваемой специальности слова. При усечении первого файла до 231 слова выполним сравнение с резюме «инженер РЭА», «инженер швейного производства» и «инженер-технолог в нефтедобывающей отрасли» (табл. 1).

Для первого соискателя:

```
Min word length in COMMON => 0
Read pages.....
Success comparing! See onlyone,onlytwo and COMMON files
Files:
[treb2.txt]=231 words [rezrea.txt]=235 words All=466
[onlyone]=203 [onlytwo]=207 [common]=28 All=466
Files metrics is correct
1-st Equal metric = 0.120181 [12%] ->Hihg
Null-Equal metric = 0.120172 [12%] ->Medium
2-d  Equal metric = 0.063927 [6%] ->Down
Medium = 0.101325 [10%]
```

Для второго соискателя:

```
Min word length in COMMON => 0
Read pages.....
Success comparing! See onlyone,onlytwo and COMMON files
Files:
[treb2.txt]=231 words [rt4.txt]=315 words All=546
[onlyone]=207 [onlytwo]=291 [common]=24 All=546
Files metrics is correct
1-st Equal metric = 0.090043 [9%] ->Hihg
Null-Equal metric = 0.087912 [8%] ->Medium
2-d  Equal metric = 0.045977 [4%] ->Down
Medium = 0.074569 [7%]
```

Таблица 1

Сопоставительный анализ вакансии и группы резюме с помощью комплекса индексирования и анализа текстов (КИАТ), усечение до 231 слова

Вакансия	Резюме	Среднее значение от 3-х метрик
Инженер-конструктор РЭА	Инженер РЭА	0,101325 [10%]
Инженер-конструктор РЭА	Инженер швейного производства	0,074569 [7%]
Инженер-конструктор РЭА	Инженер-технолог в нефтедобывающей отрасли	0,032651 [3%]

Для третьего соискателя:
Min word length in COMMON => 0
Read pages.....
Success comparing! See onlyone,onlytwo and COMMON files
Files:
[treb2.txt]=231 words [rt2.txt]=181 words All=412
[onlyone]=223 [onlytwo]=173 [common]=8 All=412
Files metrics is correct
1-st Equal metric = 0.039415 [3%] ->Hihg
Null-Equal metric = 0.038835 [3%] ->Medium
2-d Equal metric = 0.019802 [1%] ->Down
Medium = 0.032651 [3%]

Таким образом, при усечении файла требований до ключевых слов по специальности результат сравнения становится весьма убедительным, а при усечении словаря первого файла (вакансии) до ключевых слов – терминов, характеризующих компетенции сотрудника, предлагаемая нами методика вполне может быть применима для отбора объявлений с резюме, подходящими для соответствующих вакансий.

СЕМАНТИЧЕСКОЕ ЯДРО КАК ОСНОВА ФОРМУЛИРОВАНИЯ СПИСКА КЛЮЧЕВЫХ НАВЫКОВ

Любой рынок труда можно отнести к динамически развивающимся структурам, и это значительным образом влияет на выбор актуальной базы данных. Среди исследователей обозначилась заметная тенденция искать обновленные данные в интернет-ресурсах и социальных сетях. Пополняемые в режиме реального времени источники информации сами уже в течение многих лет являются предметом специального исследования. Для работы с большими интернет-данными создаются правила и специальные методики. Д. Левандовски в 2012 г. закрепил основное правило информационного поиска при работе с большими массивами открытых данных (informational retrieval), которое, в частности, заключалось в создании стабильных файлов с данными [15]. Для Д. Левандовски это правило имело смысл в контексте анализа пользовательских запросов при поиске информации в сетях или библиотечных каталогах, подключенных к открытым источникам информации. На самом же деле оно распространяется на любые виды информационного поиска, связанные с большими данными в WWW.

Для обзора возможностей семантического анализа вакансий при выделении ключевых навыков по определенной специальности рассмотрим объявления о работе по специальности «инженер-конструктор». Если речь идет не о мягких навыках (например, умении работать с документами), а об узкопрофессиональных компетенциях, то эксперту выделить их в объявлении достаточно просто (как правило, эти компетенции перечисляются в рубрике «требования»). При работе с объявлениями были выделены 7 терминов, употребленных в объявлениях более одного раза. Для этой специальности уровень профессионализма определяется набором специализированных программных продуктов, которыми владеет соискатель. Термины, имеющие варианты написания, были приведены к единообразию для удобства дальнейшей работы (табл. 2).

Таблица 2

Ранжирование терминов по частоте упоминания специальность (инженер-конструктор)

Наименование термина	Частота упоминания
AutoCad	14
SolidWorks	9
Компас-3D	5
Scad	3
Lira	3
Revit	2
Inventor	2

При анализе ключевых терминов в любом тексте большое значение имеет не только частота употребления терминов, но и поиск терминов-спутников, которые сами по себе могут редко употребляться, однако всегда сопровождают ключевые термины с большим числом употреблений. Для выделения этих «спутников» составим общую схему совместного употребления терминов (одна строка соответствует перечню терминов, упомянутых в одном объявлении):
AutoCad, SolidWorks, Revit, Lira
AutoCad, SolidWorks
SolidWorks
AutoCad, Компас-3D, Inventor, SolidWorks
AutoCad, Компас-3D, SolidWorks

Таблица 3

Совместное употребление терминов с учетом «близости» между ними

№ пары	Термин 1	Термин 2	Совместное употребление (близость)
1	AutoCad	SolidWorks	8 (0,3)
2	AutoCad	Revit	1 (0,8)
3	AutoCad	Lira	3 (0,6)
4	AutoCad	Компас-3D	4 (0,5)
5	Компас-3D	Inventor	2 **(0,4)**
6	SolidWorks	Lira	1 (0,8)
7	Revit	Lira	1 (0,6)
8	AutoCad	Inventor	2 (0,7)
9	AutoCad	Scad	2 (0,7)
10	SolidWorks	Компас-3D	4 (0,4)
11	Scad	Revit	1 (0,6)
12	Scad	Lira	1 (0,6)

AutoCad, Scad
AutoCad
SolidWorks, Компас-3D, AutoCad.
SolidWorks, AutoCad
AutoCad
Scad, Revit
SolidWorks, Компас-3D
AutoCad, Lira
AutoCad, Inventor, Компас-3D
AutoCad, SolidWorks
AutoCad, Scad, Lira
SolidWorks, AutoCad

Рассчитаем для каждой пары терминов близость их взаимного расположения по формуле, учитывающей количество совместного употребления двух терминов, а также упоминание каждого из терминов за вычетом числа совместных употреблений. Близость между терминами 1 и 2 определяется по формуле[1]:

$$(А+Б)/(2С+А+Б),$$

где: С – [Совместное употребление термина 1 и 2];
А – [Упоминание термина 1 минус С];
Б – [Упоминание термина 2 минус С].

Результатом этого шага станет цифровой показатель близости между двумя терминами (табл. 3).

Пороговые показатели значимой близости терминов могут определяться в зависимости от разницы показателей близости и целей исследования. В нашем случае определим учитываемый порог близости от 0 до 0,5 и проанализируем термины, близкие трем наиболее употребимым терминам: AutoCad, SolidWorks и Компас-3D (см. табл. 2). Анализ близости терминов позволяет обратить внимание на термин Inventor. В случае если бы пороговая значимость определялась в диапазоне от 0 до 0,6, то значение приобрел бы также термин Lira.

Таким образом, с помощью выделения семантического ядра объявлений о работе и дальнейшего статистического анализа терминов возможно создать структуру ключевых навыков для конкретной специальности, что имеет огромное значение для дальнейшей аналитики рынка труда.

ЗАКЛЮЧЕНИЕ

Рассмотренная методика выделения семантического ядра текстового массива позволит составить стабильную базу данных вакансий и соответствующую ей базу данных резюме. Автоматическое сопоставление текстов будет результативным, если его выполнять на основе определенной экспертным образом базовой семантической структуры анализируемого теста.

СПИСОК ЛИТЕРАТУРЫ

1. Анисимова А.Э., Гагельстром А.О. Российское высшее образование и рынок труда // Россия: Тенденции и перспективы развития. – 2016. – № 3. – С. 717–726.
2. Kun Lu, Xin Cai, Ajiferuke I., Wolfram D. Vocabulary size and its effect on topic representation // Information processing and management. – 2017. – Vol. 53. – P. 653–665. – URL: http://dx.doi.org/10.1016/j.ipm.2017.01.003
3. Blei D., Ng A.Y., Jordan M.J. Latent Dirichlet allocation // Journal of machine learning research. – MIT press, 2003. – Vol. 3. – P. 993–1022.
4. Hassan S.U., Haddawy P. Analyzing knowledge flows of scientific literature through semantic links: A case study in the field of energy // Scientometrics. – 2015. – Vol. 103, № 1. – P. 33 – 46.
5. Hu J., Zhang Y. Research patterns and trends of recommendation system in China using co-word analysis // Information processing & management. –

[1] В данном случае применяется мера Ланса и Уильямса, принимающая значение от 0 и 1 и учитывающая только те наблюдения, для которых присутствует хотя бы один признак.

2015. – Vol. 51, № 4. – 329–339. – URL: http://dx.doi.org/10.1016/j.ipm.2015.02.002
6. Khasseh A.A., Soheili F., Moghaddam H.S. Intellectual structure of knowledge in iMetrics: A co-word analysis // Information processing and management. – 2017. – Vol. 53. – P. 705-720. – URL: Mode of access: http://dx.doi.org/10.1016/j.ipm.2017.02.001
7. Ronda-Pupo G.A., Guerras-Martin L.A. Dynamics of the evolution of the strategy concept 1962–2008: A co-word analysis // Strategic management journal. – 2012. – Vol. 33, № 2. – P. 162–188. – DOI: 10.1002/smj
8. Gottipati S., Shankararaman V. Competency analytics tool: Analyzing curriculum using course competencies // Education and information technologies. – Springer, 2017 (published online 22 feb.). – DOI: 10.1007/s10639-017-9584-3
9. Ducrot J., Miller S., Goodman P.S. Learning outcomes for a business information systems undergraduate program // Communications of the Association for information systems. – 2008. – Vol. 23. – Article 6. – URL: http://repository.cmu.edu/cgi/viewcontent.cgi?article=1920&context=tepper
10. GnanaSingh A.A., Leaveline E.J. Competency-based calisthenics of learning outcomes for engineering education // International journal of education and learning. – Science and engineering research support society, 2013. – Vol. 2, № 1. – P. 25–34.
11. Kuo-Yu Huang, Yea-Ru Chuang. Aggregated model of ttf with utaut2 in an employment website context // Journal of data science. – 2017. – Vol. 15. – P. 187–204.
12. Googhue D.L., Thompson R.L. Task-technology fit and individual performances // MIS quarterly. – 1995. – Vol. 19, № 2. – P. 213–236.
13. Drosou M., Jagadish H.V., Pitoura E., Stoyanovich J. Diversity in big data: A review // Big data. – Mary Ann Liebert, 2017. – Vol. 5, № 2. – P. 73 – 84. – DOI: 10.1089/big.2016.0054
14. Рязанова А.А., Щербаков А.Ю. К вопросу о метриках сходства тестов для методов их автоматизированного сравнения // Технические науки: научные приоритеты ученых. Сб. научн. тр. по итогам международной научно-практической конференции. Вып. 1. – Тольятти, 2017.
15. Lewandowski D. A framework for evaluating the retrieval effectiveness of search engines // Next generation search engines / eds. C. Louis, I. Biskri, J.-G. Ganascia, M. Roux. – Hershey, PA: IGI global, 2012. – P. 456–479. – DOI: 10.4018/978-1-4666-0330-1.ch020

Материал поступил в редакцию 29.08.17.

Сведения об авторах

АНИСИМОВА Алина Эмануиловна – кандидат культурологии, ученый секретарь ВИНИТИ РАН, Москва
e-mail: uchsekr@viniti.ru

РЯЗАНОВА Алина Александровна – научный сотрудник ВИНИТИ РАН, Москва
e-mail: vvshpp@mail.ru

ЩЕРБАКОВ Андрей Юрьевич – доктор технических наук, профессор НИУ ВШЭ, главный научный сотрудник Федерального исследовательского центра «Информатика и управление» РАН, Москва
e-mail: x509@ras.ru

УДК 022.9 : 004.9

И.В. Тимошенко

Технология радиочастотной идентификации в библиотеках

Обсуждаются основные направления развития технологии радиочастотной идентификации (РЧИ) и проблемы стандартизации применительно к библиотечным технологиям. Предлагаются подходы к созданию библиотечных систем РЧИ, интегрированных в глобальную сеть электронного кода продукта на основе гармонизированной нормативной базы.

Ключевые слова: *радиочастотный идентификатор, библиотечная система, стандартизация, информационные технологии*

Технология радиочастотной идентификации (РЧИ) прочно вошла во многие области деятельности и жизни современного информационного общества. Сегодня актуальной задачей является осмысление полученного опыта, закрепление его в виде общепринятых правил, выполнение которых может обеспечить дальнейшее развитие как прикладных технологий, связанных с РЧИ, так и самой технологии радиочастотной идентификации.

Хотя основы радиочастотной идентификации закладывались ещё в 30-40-е гг. XX в., но только в 1990-е г. началось её бурное развитие, что было обусловлено успехами в развитии цифровой техники и микроэлектроники. Начало развития технологии радиочастотной идентификации в современном её понимании было положено в Массачусетском технологическом институте (США), где впервые были разработаны стандарты, необходимые для широкого применения этой технологии на практике. Стандартные подходы позволили наладить массовое производство меток и сделать их доступными во многих областях. Финансовую поддержку этого проекта оказывала организация *Uniform code council, inc.*[1]. В 1999 г. в рамках проекта был открыт специализированный научный центр *Auto-ID center* в Кембриджском исследовательском центре Массачусетского технологического института, а также аналогичные центры при университетах Австралии, Великобритании, Китая, Кореи, Швейцарии, Японии. В 2003 г. *Auto-ID center* был преобразован в научное объединение *Auto-ID labs*, которое совместно с созданной организацией *EPC global* продолжило развитие и стандартизацию технологии радиочастотной идентификации.

Еще одно направление в области стандартизации средств РЧИ связано с работой таких международных организаций, как International Organization for Standardization (ИСО) и International Electrotechnical Commission (МЭК). В 1987 г. ими был образован Совместный технический комитет (СТК) ИСО/МЭК СТК 1 «Информационные технологии», в рамках которого в 1996 г. был создан Подкомитет ISO/IEC JTC1/SC31 Automatic identification and data capture techniques ПК 31 «Автоматическая идентификация и технология сбора данных», в котором были разработаны первые стандарты, упорядочивающие технические характеристики различных устройств радиочастотной идентификации и методы их применения.

Сегодня состояние стандартизации оборудования РЧИ представляет собой достаточно разнообразную картину, отраженную в ряде стандартов ИСО/МЭК и *GS1/EPC global*, определяющих разные типы оборудования, работающего в разных частотных диапазонах и по разным протоколам.

Базовым стандартом СТК ИСО/МЭК для всех видов устройств РЧИ сегодня является ИСО/МЭК 18000 под общим названием «Информационные технологии. Радиочастотная идентификация для управления предметами». Стандарт получил широкую, хотя и не всеобщую поддержку со стороны производителей оборудования РЧИ. Этот стандарт представлен шестью частями, определяющими параметры радиоинтерфейса устройств радиочастотной идентификации, работающих во всех установленных диапазонах частот.

Кроме стандартов ИСО/МЭК, широкое распространение в мире получили стандарты *EPC global* – организации, продвигающей концепцию единого Электронного кода продукта, как единого идентификатора для всех систем РЧИ. Организацией была предложена классификация устройств РЧИ относительно их функциональных возможностей. На сегодняшний день существует два поколения стандартов *EPC global* (*Electronic product code*). Стандарты первого поколения – *Generation 1* включают специфика-

[1] Некоммерческая организация, занимающаяся установлением и продвижением стандартов идентификации продукции и соответствующих средств электронных коммуникаций в США. В 2005 г. вошла в глобальную организацию по стандартизации GS1.

ции для меток Класса 0 и Класса 1. Эти стандарты были разработаны организацией *Auto-ID labs* и обнародованы вместе с появлением *EPC global*. Представленные в них метки имели несовместимые протоколы обмена данными. При разработке новой версии стандартов – *Generation 2* был разработан новый протокол Класса 1 (*EPCC1g2*), заместивший протоколы Классов 0 и 1 предыдущего поколения. В стандартах *Generation 2* были определены функциональные отличия для меток Классов 2, 3 и 4. Изначально стандарты *EPC global* были разработаны только для оборудования СВЧ диапазона (850–960 МГц) как наиболее востребованного в области складской и транспортной логистики.

Существование двух различных групп стандартов, определявших работу сходных типов устройств и не совместимых между собой, было существенным препятствием для развития систем РЧИ. Ни одна из них не была полностью поддержана производителями оборудования. В двух диапазонах, наиболее используемых в практических областях − ВЧ (13,56 МГц) и СВЧ (850−960 МГц), ведущие производители ВЧ оборудования РЧИ (в том числе библиотечного) стали использовать стандарт ИСО/МЭК *18000-3Mode* 1 (*ISO/IEC 18000-63:2015 Information technology – Radio frequency identification for item management. – Part 63: Parameters for air interface communications at 860 MHz to 960 MHz)*, а производители СВЧ оборудования – *EPC C1g2 (EPC™ UHF Class 1 Generation 2 air interface specification (Release 1.02)*). Применение оборудования того или иного диапазона в конкретных областях определялось их характеристиками и ограничениями, вытекающими из физических свойств электромагнитных волн. Кроме того, логические устройства меток этих диапазонов существенно отличались друг от друга, что наряду с различием частотных диапазонов, делало ВЧ и СВЧ системы радиочастотной идентификации альтернативными друг другу.

Первый шаг в направлении гармонизации двух направлений стандартизации был сделан СТК ИСО/МЭК. В 2006 г. появилось дополнение к существующей группе стандартов ИСО/МЭК 18000, был принят стандарт ИСО/МЭК 18000-63 *(ISO/IEC 18000-3:2010 Information technology – Radio frequency identification for item management. – Part 3: Parameters for air interface communications at 13,56 MHz)*, определяющий протокол обмена данными между устройствами радиочастотной идентификации СВЧ диапазона, совместимый с протоколом типа *EPCC1g2*.

Следующим шагом можно считать разработку в 2011 г. международными организациями *EPC global* совместно с GS1 стандарта *EPC Class 1 HF (EPC™ Radio-frequency identity protocols EPC Class-1 HF RFID Air interface protocol for communications at 13.56 MHz Version 2.0.3. – 5 september, 2011),* определяющего протоколы концепции EPC для оборудования высокочастотного диапазона. Появление нового стандарта было поддержано ИСО/МЭК СТК1/ПК31 принятием аналогичного дополнения *Mode 3* к стандарту ИСО/МЭК 18000-3.

Появление общих стандартных подходов к производству и применению оборудования РЧИ в наиболее востребованных частотных диапазонах создало принципиальную возможность реализации изначальной концепции *EPC global* об использовании единого электронного кода продукта (*Earning Per Click*) для идентификации объектов учета в системах РЧИ различной специализации, включая библиотечные. Реализация такой идеи представляется актуальной, так как сегодня технология радиочастотной идентификации приобретает черты всеобщей технологии. Устройства радиочастотной идентификации широко используются или планируются к применению во многих областях. Именно с этой технологией связывается развитие концепции «Интернет вещей» (Internet of things), возникшей в середине 2000-х гг. и предполагающей активное присутствие вещей, окружающих человека, в глобальной сети Интернет, а также появление средств коммуникации как между людьми и вещами, так и между самими вещами непосредственно. Технология РЧИ активно внедряется в промышленности, в транспортной и складской логистике, в магазинах, медицине и в сельском хозяйстве, а также в библиотеках.

Широкое внедрение технологии РЧИ в библиотеках началось в начале XXI в. В 2005 г. в Дании был принят первый национальный стандарт, определяющий правила применения радиочастотных идентификаторов в библиотеках, который был поддержан во многих странах. В 2011 г. техническим комитетом ИСО ТК46/ПК4 был принят международный стандарт ИСО 28560 «Информация и документация – Радиочастотная идентификация в библиотеках», в котором были регламентированы основные технические параметры библиотечных систем РЧИ, а также структуры и протоколы обмена данными с библиотечными системами автоматизации. Стандарт ИСО 28560 представляет собой группу стандартов и состоит из четырёх частей:

часть 1: Элементы данных и общее руководство по применению;

часть 2: Кодирование элементов данных РЧИ на основе правил стандарта ИСО/МЭК 15962;

часть 3. Кодирование фиксированной длины;

часть 4. Кодирование элементов данных для радиочастотной идентификации, основанных на правилах ИСО/МЭК 15962, в радиочастотных метках с разделенной памятью.

В первой части стандарта определены элементы данных, используемые при каталогизации документов библиотечного фонда, которые могут быть размещены в памяти меток РЧИ и использованы для автоматизации технологических операций в библиотеках. Всего задано 26 таких элементов, из них обязательными являются только два: первичный идентификатор документа и стандартный идентификационный код библиотеки (ISIL). Элементы данных приводятся без указания условий их размещения в памяти меток, которая в общем случае может иметь различную организацию для разных типов меток.

Вторая часть стандарта устанавливает способ размещения в памяти меток элементов данных, определенных в первой части, основанный на стандартных правилах кодирования структуры идентификаторов объекта, определенных в стандарте ИСО/МЭК

15962. Их применение позволяет наиболее рационально использовать имеющуюся память меток РЧИ, без привязки к конкретному типу меток.

Третья часть стандарта основана на датском национальном стандарте (часто называемом «Датской моделью данных») и опыте его применения в других странах. Принципы размещения элементов данных, установленные в стандарте, основаны на фиксированной блочной структуре данных, состоящей из полей фиксированной и переменной длины. Принципы размещения данных, определенные в этой части стандарта, не совместимы с правилами, изложенными во второй части, и носят более жесткий характер. Представленные в стандарте структуры данных ориентированы, прежде всего, на метки ВЧ диапазона, соответствующие ГОСТ Р ИСО/МЭК 18000-3 *Mode1*. Остальные типы меток рассматриваются в этой части стандарта с точки зрения степени их совместимости с базовым типом.

В целом кодирование данных, основанное на правилах третьей части стандарта, менее рационально в сравнении с правилами, представленными во второй его части. Принятие этого стандарта объясняется тем, что кодирование, основанное на правилах Датской модели, стало для библиотек «де-факто» международным стандартом задолго до того как такой стандарт был принят Совместным техническим комитетом ИСО/МЭК. Большое количество библиотек во многих странах мира имеют огромное количество документов, маркированных метками радиочастотных идентификаторов ВЧ диапазона, закодированными по этим правилам. Переход на другие типы меток и методы кодирования – это и по настоящее время практически трудно реализуемая задача. Из других областей, где широко используются метки указанного типа, можно назвать прокат спортивного инвентаря и маркировку животных. В других областях такие метки практически не используются.

Четвертая часть стандарта появилась позднее трёх предыдущих и была принята только в 2014 г. Здесь определены правила размещения элементов данных, представленных в первой части стандарта, согласованные с правилами кодирования, определёнными во второй его части. Представленные в стандарте структуры данных ориентированы на метки радиочастотных идентификаторов, имеющие блочную организацию памяти, установленную в стандарте *EPC global* как Класс 1 *Generation 2* (*EPCC1g2*). Стандарт ориентирован на метки СВЧ диапазона (860−960 МГц), что формально сужает его область действия, поскольку фактически описанную в стандарте структуру памяти в настоящее время могут реализовать метки радиочастотных идентификаторов двух типов:

- ВЧ − ГОСТ Р ИСО/МЭК 18000-3 Mode1; *EPC Class 1 HF*;
- СВЧ − ГОСТ Р ИСО/МЭК 18000-6 тип С; *EPCC1g2*.

Положения, приведенные в стандарте, принципиально могут быть отнесены к обоим типам меток в равной степени. При таком подходе появление четвертой части стандарта можно считать важным шагом в направлении гармонизации локальных систем радиочастотных идентификаторов обоих частотных диапазонов, что вписывается в генеральное направление развития технологии РЧИ.

В целом можно сказать, что четвертая часть стандарта определяет правила кодирования меток, применимые для меток, имеющих структуру памяти, соответствующую спецификации EPC *Class 1 Generation 2*, при этом структура уникального идентификатора предмета учёта, размещаемого в области памяти EPC, несовместима с форматом кода EPC. Это обстоятельство делает библиотечные системы радиочастотных идентификаторов альтернативными EPC системам. Следует отметить, что системы РЧИ, основанные на правилах, определенных в четвёртой части стандарта, полностью совместимы по элементам данных с «традиционными» системами, применяемыми в большинстве библиотек мира, определёнными в третьей части стандарта и основанными на «Датской модели данных».

Использование подходов, изложенных в четвёртой части стандарта, дает принципиальную возможность создания «расширенной» библиотечной автоматизированной системы РЧИ, имеющей унифицированный набор функций для работы с метками двух конкурирующих в настоящее время диапазонов частот: ВЧ и СВЧ, гармонизированных с функциональностью существующих библиотечных высокочастотных систем и радиочастотных идентификаторов.

«Прозрачная» работа системы РЧИ в двух диапазонах, наряду с использованием EPC меток, требует изменения двучастотных считывателей радиочастотных идентификаторов. Создание таких считывателей представляет собой технически сложную задачу. Пробный шаг в этом направлении был сделан компанией *FEIG electronic*, которая в 2013г. начала производство мобильных ридеров типа *IDISC.PRHD102*, поддерживающих одновременную работу в ВЧ/СВЧ диапазонах.

Несмотря на наличие считывателя радиочастотного идентификатора, поддерживающего работу в двух диапазонах, его использование в предлагаемой «расширенной» системе на сегодняшний день невозможно, так как в нём не реализована поддержка работы с ВЧ метками типа ИСО/МЭК 18000-3 *Mode 3* (*EPC Class 1 HF*).

Возможность производства EPC меток ВЧ диапазона появилась 2013 г., когда компания NXP начала выпуск чипов типа *ICODE ILT*, которые соответствуют стандарту ИСО/МЭК 18000-3 *Mode 3 (EPC Class 1 HF*). На базе этих чипов возможно производство библиотечных EPC меток ВЧ диапазона, но до сегодняшнего дня такие метки на рынке не представлены и не используются в системах радиочастотных идентификаторов.

Таким образом, появление «расширенных» библиотечных систем радиочастотных идентификаторов на сегодняшний день проблематично. Разработчики систем радиочастотных идентификаторов сталкиваются с проблемой отсутствия на рынке необходимого оборудования – меток и считывателей, а производители не спешат вкладывать средства в налаживание производства нового оборудования из-за несформировавшегося рынка сбыта. Ситуация похожа на ту,

которая складывалась в конце 1990-х г. с технологией радиочастотной идентификации в целом. Изменить положение может реализация крупного проекта системы РЧИ с использованием EPC меток ВЧ диапазона в области, где применение меток этого диапазона целесообразно, наряду с метками СВЧ диапазона. В рамках такого проекта могут быть сделаны наработки, дальнейшая коммерциализация которых изменит ситуацию на рынке. Реализация такого проекта на уровне, обеспечивающем его экономическую эффективность, под силу только крупной коммерческой или государственной организации. Библиотечные системы радиочастотной идентификации занимают очень скромное место в общем объеме РЧИ систем и вряд ли смогут обеспечить своими потребностями проект требуемого масштаба.

Нужный масштаб может иметь проект, реализующий давно обсуждаемую идею маркировки печатных изданий на стадии их производства с использованием информации, записанной на метке, на всех этапах в цепи поставок от типографии до магазина или библиотеки с последующим использованием её при каталогизации и работе с библиотечным фондом. Реализация такого проекта возможна при условии включения библиотечных систем РЧИ в систему Глобальной сети *EPC (EPC global Network)*. Принципиальная возможность реализации такого проекта вытекает из гармонизации EPC стандартов для ВЧ и СВЧ меток, а также из наличия библиотечного стандарта для EPC меток, гармонизированного с Датской моделью данных. Применение в таком проекте ВЧ меток предпочтительнее по сравнению с СВЧ метками, исходя из физических свойств электромагнитного поля используемых диапазонов частот, при этом маркированные издания могут попадать в сферу действия автоматизированных систем радиочастотных идентификаторов, работающих по правилам *EPC*, при транспортировке и складировании.

Реализация такого проекта потребует применения двух частотного оборудования, работа которого основана на гармонизированной нормативной базе, прозрачно идентифицирующей метки ВЧ и СВЧ диапазонов.

Как ещё одну проблему, требующую решения в рамках реализации такого проекта, можно рассматривать несовместимость принципов формирования уникального идентификатора в библиотечных системах и кода EPC. Отсутствие такой совместимости делает библиотечные системы автоматизации «не вхожими» в Глобальную сеть EPC, что препятствует интеграции библиотечных технологий в смежные технологические области, связанные с радиочастотной идентификацией. Принципиально такая возможность существует, учитывая технические характеристики меток, применимых в библиотечных системах РЧИ, но реализация этой возможности связана с модификацией существующего «библиотечного» стандарта ГОСТ Р ИСО 28560 и требует расширения нормативной базы технологии *EPC*.

Электронный код продукта – *EPC*, предлагаемый организациями *EPC global* совместно с *GS1* представляет собой числовой идентификатор, уникальный для каждого материального объекта, подлежащего учету. В настоящее время наиболее часто используются коды стандартной длины 64 и 96 бит. Существуют и планируются к внедрению также 128 и 256 битовые коды. Общая длина кода определяет возможную длину полей данных, а также как следствие ширину кодового пространства и свободу выбора форматов представления данных. Базовая структура кода *EPC* состоит из четырех полей, назначение которых установлено следующим образом:

- заголовок определяет тип кода;
- номер поставщика/владельца определяет организацию − владельца объекта учета;
- класс объекта определяет типовую идентичность объекта учета;
- серийный номер идентифицирует конкретный экземпляр объекта.

Поля данных представлены в коде как жесткая структура полей фиксированной длины, размер которых определяется типом кода. Значения первых двух полей регламентируются и устанавливаются стандартами *EPC global* и *GS1*. Последние два поля устанавливаются организацией-владельцем исходя из локальных технологических потребностей. Эти поля могут быть перезаписаны при переходе от одного логистического этапа к другому.

При поступлении в библиотеку издания, маркированного на ранних этапах цепи поставок согласно ИСО 28560-4, весь блок 01 *(EPC)* памяти его метки должен быть представлен значением уникального идентификатора предмета учета и байта *AFI*. Информация о коде *EPC* при этом теряется, а формат записанной информации не соответствует формату *EPC*, таким образом метка перестает опознаваться в *EPC* системах радиочастотных идентификаторов. В случае возврата метки в такую автоматизированную систему, например, при доставке получателю через транспортную компанию, почтовую службу или при попадании в торговую сеть, в её памяти нужно восстановить код *EPC* с потерей «библиотечной» информации.

Техническая возможность применения идентификационного кода в формате *EPC* в библиотечных автоматизированных системах существует, но на сегодняшний день она не предусмотрена библиотечными стандартами. Применительно к печатным изданиям в случае их маркировки при изготовлении в типографии на начальном этапе логистической цепочки издательство может присваивать свои значения типа издания и серийного номера, которые могут быть использованы для автоматизации технологических этапов при транспортировке и хранении изданий. При поступлении таких документов в библиотеку эти поля могут быть переназначены в соответствии с правилами каталогизации библиотечного фонда, при этом значение поля «номер поставщика/владельца» может быть использовано при каталогизации как идентификатор издательства или организации – поставщика документа. Суммарная длина переназначаемых полей (в рамках формата *EPC*) может быть от 60 до 160 бит. Из них поле *класс объекта* длиной 13−24 бита может быть использовано для размещения идентификатора вида издания *RUSMARC* или *ONIX*, который может быть общим для книгоиздате-

лей и библиотек. В этом случае перезаписи может подлежать только поле «серийный номер», под которое, для разных типов *EPC* меток, может быть отведено от 36 до 180 бит. В этом поле могут содержаться элементы данных, составляющие уникальный идентификатор предмета учёта (*UII*) согласно ИСО 28560-4. Общая длина *UII*, согласно Датской модели данных, составляет 19 байт и складывается из следующих компонент:

- первичный идентификатор − 16 байт,
- код ISIL − 11 байт,
- информация о комплекте − 2 байта.

При кодировании по правилам *URNCode40* общая длина кода составит 10 байт. Таким образом, общая длина записи в поле «серийный номер» вместе с добавленным значением байта семейства приложений *AFI* составит 88 бит. Полученный размер поля не превышает максимального возможного размера для формата *EPC* и полностью вмещается в составе полного кода *EPC* в блок 01 меток типа *NXPICODEILT*, размер которого составляет 240 бит.

Приведенный пример показывает только принципиальную возможность гармонизации стандартных библиотечных систем автоматизации и систем радиочастотной идентификации, основанных на стандартах *EPC*. Для реализации такой возможности потребуется модификация нормативной базы как библиотечного стандарта ИСО 28560 на уровне технического комитета ИСО ТК46/ПК4, так и правил формирования − *EPC* (Электронного кода продукта) на уровне организаций *EPC global* и *GS1*.

Возможность реализации проекта, в рамках которого будут решены такие задачи, на сегодняшний день назрела и вытекает из общей логики развития информационных систем и, в частности, технологии радиочастотной идентификации. Для логистических *EPC* систем возможность прозрачного использования меток ВЧ и СВЧ диапазонов позволит эффективно использовать маркировку метками радиочастотной идентификации предметов учета в областях, где метки ВЧ диапазона наиболее предпочтительны в силу физических свойств электромагнитных волн ВЧ диапазона. К таким областям можно отнести, прежде всего, издательскую деятельность, библиотеки, а также производство жидких лекарств, парфюмерии, различных напитков. Для библиотек реализация такого проекта позволит кардинально повысить доступность технологии радиочастотных идентификаторов для библиотечных автоматизированных систем за счет комплектования фонда печатными документами, уже маркированными в издательстве метками совместимого формата. Включение маркированных документов библиотечного фонда в информационное пространство Глобальной сети *EPC* может в перспективе существенно повысить их мобильность в службах доставки системы МБА, а также доступность фонда для читателей за счет широкого использования новых информационных технологий с использованием глобальной идентификации. В целом можно утверждать, что использование технологии радиочастотных идентификаторов в библиотеках, основанное на нормативной базе, гармонизированной с глобальными технологиями идентификации, существенно повысит эффективность интеграции документов традиционного библиотечного фонда наряду с электронными документами в современное информационное пространство в рамках новых концепций развития библиотечных технологий.

Материал поступил в редакцию 29.08.17.

Сведения об авторе

ТИМОШЕНКО Игорь Владимирович – кандидат технических наук, ведущий научный сотрудник, Главный технолог автоматизированных систем ГПНТБ России, Москва
e-mail: timigor@gpntb.ru

ДОКУМЕНТАЛЬНЫЕ ИСТОЧНИКИ ИНФОРМАЦИИ

УДК 004.65 : [002 : (54+66)]

Л.М. Королева, Е.В. Колтунова

О представлении химической информации в реферативных базах данных

Рассмотрены некоторые вопросы систематизации научных публикаций по химии и химической технологии в реферативных базах данных Web of Science Core Collection, Scopus и БД Химия ВИНИТИ РАН. Отмечен бурный рост носящих междисциплинарный характер публикаций ученых-химиков. Обсуждаются проблемы информационного обеспечения, связанные с усилением меж- и мультидисциплирного характера исследований. Приведены результаты поиска информации в реферативных базах данных по основным предметным областям. Показано увеличение темпов роста научных публикаций по химии и химической технологии за последние годы по сравнению с другими отраслями знания.

Ключевые слова: *научная публикация, химическая информация, реферативные базы данных, публикационная активность, информационное обеспечение, предметная область, Web of Science Core Collection, Scopus, БД Химия ВИНИТИ РАН*

Химическая наука является одной из трех научных областей, за достижения в которой в течение уже 117 лет, ежегодно, начиная с 1901 г., Шведская королевская академия наук совместно с Нобелевскими комитетами по физике и по химии, выдвигает и затем награждает Нобелевской премией выдающихся ученых – тех, кто, согласно завещанию Альфреда Нобеля, написанному 27 ноября 1895 г. в Париже, «сделает наиболее важное открытие или усовершенствование в области химии...». Среди лауреатов нобелевской премии по химии – выдающийся русский химик Николай Николаевич Семенов.

Присуждаемая ежегодно самая уважаемая в мировом научном сообществе премия подчеркивает на протяжении более 100 лет значимость и ценность химической науки для жизни человеческого общества. Соответственно, невозможно переоценить значимость и ценность накопления и осмысления информации в области химии.

В сегодняшней науке все большее место занимают меж- и мультидисциплинарные исследования. Такие исследования на стыке наук, требующие участия специалистов порой из совершенно разных предметных областей, начали интенсивно развиваться в 80-е гг. прошлого столетия. Следствием бурного роста междисциплинарных проектов стало появление и формирование новых самостоятельных научных направлений. В области естественных наук появились такие мультидисциплинарные и уже самостоятельные направления, как компьютерная химия, нанотехнологии, хемоинформатика, биофизика, биоинформатика, биомедицина и т.п.

Емко и точно необходимость и неизбежность развития меж- и мультидисциплинарного подхода в науке выражает известная фраза Нобелевского лауреата по химии 2006 г. американского биохимика Роджера Корнберга: «Создание современных лекарств требует знания квантовой теории». О растущей роли и значении междисциплинарных исследований неоднократно говорил в своих интервью российский Нобелевский лауреат Жорес Алферов, подчеркивая при этом необходимость принципиально новых междисциплинарных подходов и в образовании.

Междисциплинарные исследования ведутся сегодня не только в отделениях Российской академии наук, но и в ведущих вузах страны, таких как МГУ, МФТИ, СПГУ. Один из самых широко известных междисциплинарных проектов – проект Человеческий геном *(The Human Genome Project, HGP)*, начавшийся в 1990 г., потребовал совместной работы биологов, химиков, математиков, физиков.

Такие исследования, с одной стороны, требуют соответствующего междисциплинарного информационного обеспечения, а с другой – результатом этих исследований становятся научные публикации, имеющие зачастую междисциплинарный характер. Занимаясь информацией в области химии и наук о материалах, мы наблюдаем бурный рост носящих

междисциплинарный характер публикаций ученых-химиков и, как следствие этого, сталкиваемся с проблемой совершенствования систематизации химической информации.

Реферативная, библиографическая, полнотекстовая и структурная информация в области химии и наук о материалах аккумулируется и хранится в многочисленных базах данных, среди которых, в первую очередь, надо отметить *CAS – Chemical Abstract Service* – старейшую мировую реферативную службу в области химической информации. Сегодня крупнейшие мировые «игроки» на информационном рынке – компании *Clarivate Analytics* (более 4 000 ученых и специалистов в более чем 100 странах мира) и *Elsevier* (более 7 200 ученых и специалистов в 24 странах мира) создали и поддерживают крупные информационные кластеры, в том числе в области химической информации. Представление химической информации и ее классификация, систематизация, индексирование определяют эффективность информационного обеспечения ученых и специалистов-химиков. Так, в политематической реферативно-библиографической и наукометрической базе данных *Web of Science Core Collection – WoSCC* (на платформе *Web of Science*, *Clarivate Analytics*) для систематизации знания предлагаются 5 глобальных предметных кластеров, среди которых интерес для нас представляют, в первую очередь, 3 кластера: Науки о жизни и биомедицина *(Life Sciences and Biomedicine)*, Естественные науки *(Physical Sciences)*; Технология *(Technology*). Далее каждый глобальный кластер подразделяется на предметные области, являющиеся в терминологии *Web of Science* «областями исследования» (*Research Areas*). Количество предметных областей (областей исследования) в предметных кластерах приведено в табл. 1. Во всех базах данных на платформе *Web of Science* реализован поиск информации именно по этим предметным (исследовательским) областям.

Реальным конкурентом баз данных на платформе *Web of Science* является основанный в 2004 г. агрегатор в области научной информации *Scopus* (*Elsevier*) – крупнейшая реферативная база данных и база данных научного цитирования. Основы систематизации знаний в *Scopus* следующие. Для изданий, индексируемых в этой базе данных, предлагаются 4 глобальных предметных кластера *(SubjectArea)*. Интересующая нас химическая информация попадает, в основном, в три из них: Науки о здоровье *(Health Sciences)*, Науки о жизни *(Life Sciences)*, Естественные науки *(Physical Sciences)*. Глобальные предметные кластеры разделяются, в свою очередь, на 27 основных предметных областей/разделов *(major subject areas)*, которые сформированы на основании ASJC-кодов *(All Science Journal Classification)*, и 300+ так называемых второстепенных предметных областей/разделов (minor subject areas). Поиск информации в БД *Scopus* реализован по 27 основным предметным разделам, которые в терминологии *Scopus* называются также наиболее часто используемыми *(the most frequent Subject Area categories)*. Количество предметных разделов в предметных кластерах приведено в табл. 2.

Нами был выполнен расширенный поиск *(Advanced Search)* в базах данных *WoSCC* и *Scopus* с указанием предметной области Химия *(Chemistry)* за временной период 2013-2017 гг. Отнесение найденных документов по направлениям исследования (*WoSCC*) и областям знания (*Scopus*) приведены в табл. 3 и табл. 4.

Из приведенных данных видно, что более 50% документов из предметной области «Химия» относятся к прикладным аспектам науки, таким как биофизика, молекулярная биология, компьютерные науки и т.д.

Таблица 1

Предметные кластеры в базах данных на платформе *Web of Science*

Кластер	Количество предметных областей (областей исследования)
Науки о жизни и биомедицина *(Life Sciences and Biomedicine)*	75
Естественные науки *(Physical Sciences)*	17
Технология *(Technology).*	21
Искусство и гуманитарные науки *(Arts & Humanities)*	14
Общественные науки *(Social Sciences)*	24

Таблица 2

Предметные кластеры в базе данных *Scopus*

Кластер	Количество предметных разделов (в т.ч. областей знания)
Науки о здоровье *(Health Sciences)*	103
Науки о жизни *(Life Sciences*	52
Естественные науки *(Physical Sciences)*	116
Социальные и гуманитарные науки *(Social Sciences & Humanities)*	66

Таблица 3

Отражение предметной области «Химия» в БД *WoSCC* (фрагмент)

Поле Web of Science: Направления исследований	Число записей	%
CHEMISTRY всего	918 500	100,000
в том числе:		
Materials science	167 739	18,262
Physics	126 966	13,823
Science technology other topics	89 247	9,717
Biochemistry Molecular biology	63 867	6,953
Engineering	57 304	6,239
Electrochemistry	45 009	4,900
Energy Fuels	39 207	4,269
Pharmacology Pharmacy	34 615	3,769
Food Science Technology	32 447	3,533
Polymer Science	19 809	2,157
Instruments Instrumentation	18 214	1,983
Crystallography	18 147	1,976
Metallurgy Metallurgical Engineering	17 341	1,888
Spectroscopy	13 709	1,493
Thermodynamics	12 990	1,414
Biophysics	11 250	1,225
Agriculture	10 050	1,094
Nutrition Dietetics	8 855	0,964
Biotechnology Applied Microbiology	8 787	0,957
Nuclear Science Technology	7 651	0,833
Environmental Sciences Ecology	7 491	0,816
Computer Science	5 399	0,588
Toxicology	3 776	0,411
Mathematics	3 462	0,377

Таблица 4

Отражение предметной области «Химия» в БД *Scopus*(фрагмент)

Поле Scopus: Область знания	Число записей	%
CHEMISTRY всего	1 090 185	100,000
в том числе:		
Chemical Engineering	332 867	30,533
Materials Science	327 834	30,071
Physics and Astronomy	248 946	22,835
Biochemistry, Genetics and Molecular Biology	207 586	19,041
Engineering	138 539	12,708
Medicine	72 501	6,650
Energy	66 081	6,061
Pharmacology, Toxicology and Pharmaceutics	64 573	5,923
Environmental Science	54 047	4,958
Agricultural and Biological Sciences	34 920	3,203
Computer Science	26 633	2,443
Mathematics	20 611	1,891
Social Sciences	8 216	0,754
Earth and Planetary Sciences	8 179	0,750
Immunology and Microbiology	3 602	0,330
Business, Management and Accounting	2 527	0,232
Health Professions	1 949	0,179
Arts and Humanities	955	0,088

Поле Scopus: Область знания	Число записей	%
Economics, Econometrics and Finance	194	0,018
Dentistry	159	0,015
Psychology	104	0,010
Neuroscience	53	0,005
Nursing	25	0,002
Veterinary	17	0,002

Таблица 5

Рейтинг стран по количеству документов, отраженных в предметной области «Химия» БД *WoSCC*(фрагмент)

Поле: Страны/территории	Число записей	%
Всего записей	918 500	100,000
в том числе:		
Peoples R China	248 234	27,026
USA	170 975	18,615
India	61 228	6,666
Germany	58 487	6,368
Japan	53 308	5,804
South Korea	39 715	4,324
France	39 078	4,255
England	34 463	3,752
Spain	33 619	3,660
Russia	29 267	3,186
Italy	28 338	3,085
Iran	24 604	2,679
Canada	21 160	2,304
Poland	19 563	2,130
Australia	18 551	2,020
Brazil	17 183	1,871
Taiwan	14 569	1,586
Switzerland	13 366	1,455
Netherlands	11 235	1,223
Saudi Arabia	11 008	1,198
Turkey	10 985	1,196
Sweden	10 089	1,098
Singapore	9 954	1,084
Belgium	9 321	1,015
Czech Republic	9 115	0,992

Сравнение категорий «Направления исследования» и «Область знания» в двух базах данных показывает, что, имея некоторое сходство и даже совпадения в нескольких предметных категориях, эти базы все-таки разнятся в подходе к систематизации (классификации) информации в области естественных наук.

Представляло интерес изучить отражение в предметной области «Химия» публикационной активности ученых различных стран. В табл. 5 и табл. 6 приведены результаты проведенного поиска за временной период 2013-2017 гг.

Сравнение выборок, полученных в двух базах данных, по полю «Страна/территория» показывает хорошее совпадение результатов, особенно для первых пяти стран списков.

Анализ изменения публикационной активности ученых из разных стран за период с 2008 по 2016 гг. (составлен по данным *Scopus*) (рис. 1), демонстрирует, что Китай прочно занял лидирующее положение в мире в исследованиях по химии и химической технологии, значительно опередив США.

Следует отметить, что английский язык, являющийся основным языком научных коммуникаций, начинает терять лидирующие позиции из-за широкого распространения научных публикаций на китайском языке. Этому также способствует развитие систем специализированного машинного перевода.

Таблица 6

Рейтинг стран по количеству документов, отраженных в предметной области «Химия» БД *Scopus* (фрагмент)

Поле: Страна	Число записей	%
Всего записей	1 090 185	100,000
в том числе:		
China	296 071	27,158
United States	172 187	15,794
India	75 240	6,902
Germany	72 484	6,649
Japan	63 971	5,868
United Kingdom	47 128	4,323
France	46 416	4,258
South Korea	45 609	4,184
Russian Federation	40 205	3,688
Undefined	38 508	3,532
Spain	37 199	3,412
Italy	32 423	2,974
Iran	29 508	2,707
Canada	27 205	2,495
Australia	23 293	2,137
Brazil	22 609	2,074
Poland	22 264	2,042
Taiwan	16 809	1,542
Switzerland	15 491	1,421
Saudi Arabia	13 847	1,270
Turkey	13 502	1,239
Netherlands	13 430	1,232
Sweden	11 589	1,063
Egypt	11 553	1,060
Belgium	11 210	1,028

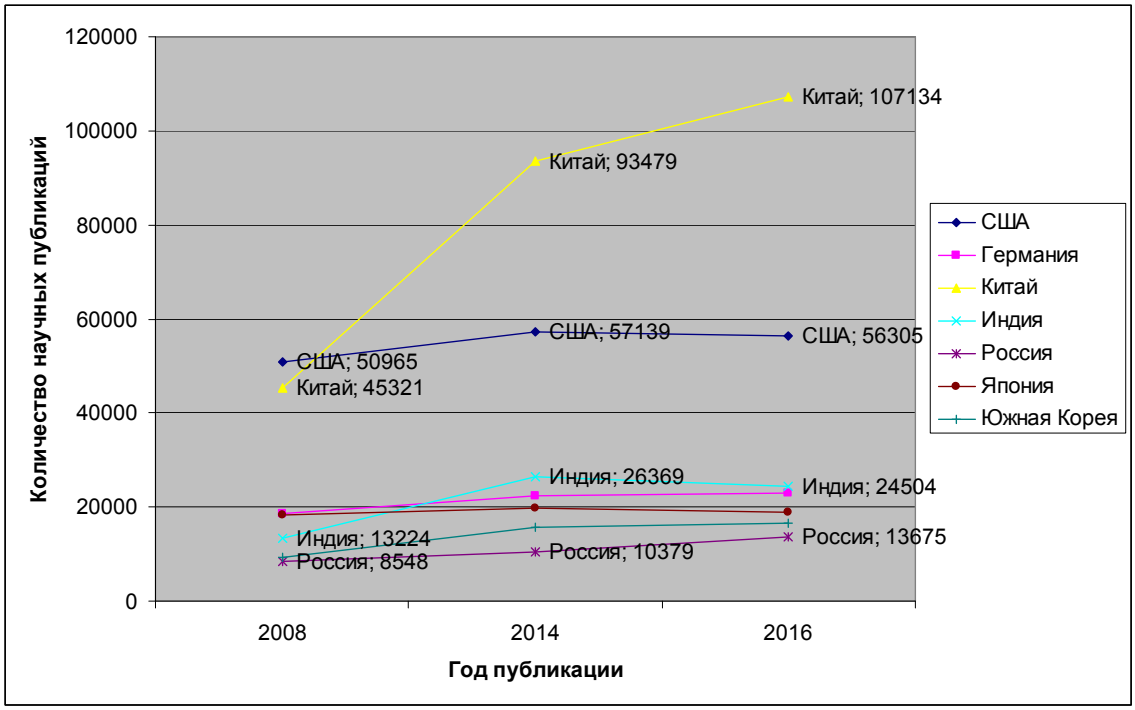


Рис.1. Динамика публикационной активности ученых некоторых стран в предметной области «Химия и химическая технология» (по данным *Scopus*)

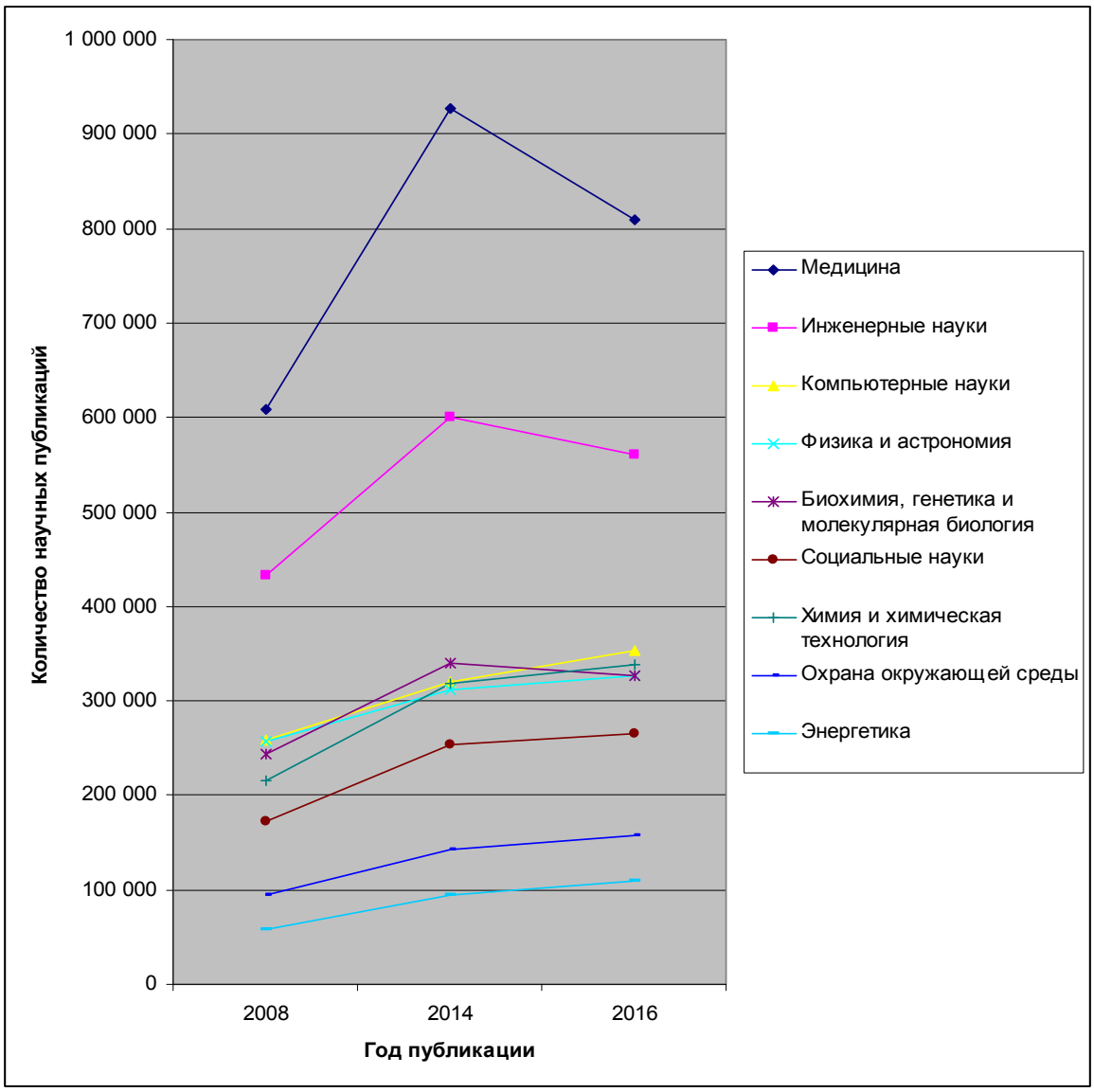


Рис. 2. Динамика изменения количества научных публикаций, посвященных исследованиям в различных предметных областях в период с 2008 по 2016 гг. (по данным *Scopus*)

По данным *Scopus* количество научных публикаций, посвященных исследованиям в области химии и химической технологии, в общем потоке научных публикаций составляет 11,8%.

За последние годы наметился постоянный рост научных публикаций по химии и химической технологии – с 216 406 публикаций в 2008 г. до 338 311 публикаций в 2016 г., т.е. прирост составил 56,3% (рис. 2).

По темпам роста научных публикаций в этот период предметная область «Химия и химическая технология» среди естественных и технических наук вышла на третье место после «Энергетики» (прирост 90,2%) и «Охраны окружающей среды» (прирост 67,4%). А по общему количеству публикаций в 2016 г. предметная область «Химия и химическая технология» переместилась на 4-е место, обогнав такие предметные области как «Биохимия, генетика и молекулярная биология» и «Физика и астрономия».

Национальная реферативная База данных в области естественных, точных и технических наук генерируется в ВИНИТИ РАН с 1953 г. в печатной форме и с 1981 г. – в печатной и электронной формах. Национальный центр информации России ВИНИТИ РАН, обладая 65-ти летним опытом поиска, сбора, обработки, использования и распространения информации, имеет значительные достижения в области систематизации и классификации научно-технической информации. Разработанный на основе углубления Государственного рубрикатора НТИ России (ГРНТИ) Рубрикатор ВИНИТИ определяет систематизацию потока научно-технической литературы до 9-го уровня. Предметные области установлены в соответствии с ГРНТИ. Рубрикатор ВИНИТИ представляет собой комплекс рубрикаторов областей знания, которые в совокупности описывают тематику научно-технической литературы, отражаемой в БД ВИНИТИ и в Реферативном журнале ВИНИТИ. База данных включает 30 тематических фрагментов (предметных

областей), состоящих из 217 разделов. Тематический фрагмент «Химия» в настоящее время включает 19 предметных разделов, соответствующих основным направлениям химической науки, каждый из 19 предметных разделов состоит из предметных рубрик 3-го уровня. Общее количество предметных рубрик – 97. Химическая информация в потоке научно-технической литературы ВИНИТИ в 2016 г. составила 25,8% от общего потока обработанных документов. В Реферативной базе данных по химии и химической технологии – самом крупном тематическом фрагменте политематической БД ВИНИТИ – осуществляется систематизация информации в области химии на основе принципа современной дифференциации наук.

СПИСОК ИСПОЛЬЗОВАННОЙ ЛИТЕРАТУРЫ

1. Web of Science Core Collection. – URL: https://clarivate.com/products/web-of-science/web-science-form/web-science-core-collection/
2. Scopus. – URL: http://www.elsevierscience.ru/products/scopus/
3. База данных ВИНИТИ РАН. – URL: http://bd.viniti.ru/index.php?option=com_content&task=view&id=238&Itemid=101
4. Рубрикатор ВИНИТИ (текущий, 2018 г.). – URL: http://scs.viniti.ru/rubtree/main.aspx?tree=RV
5. Реферативный журнал «19.Химия». Сводный том. – М.: ВИНИТИ РАН, 1953–2017 гг.

Материал поступил в редакцию 29.08.17.

Сведения об авторах

КОРОЛЕВА Любовь Михайловна – кандидат химических наук, зав. отделением научной информации по проблемам химии и наук о материалах ВИНИТИ РАН, Москва
e-mail: lkorol@viniti.ru

КОЛТУНОВА Елена Валентиновна – старший научный сотрудник отдела научной информации по проблемам химии и химической технологии ВИНИТИ РАН
e-mail: chemistry711@viniti.ru

УДК 002 : 62

М.И. Гречиков

Научно-техническая информация по машиностроению в России

Представлена история становления и развития в России научно-технической информации по машиностроению с начала XX в. и до наших дней. Отмечается, что в 20-30-х гг. XX века, с началом развития промышленности в стране, появился ряд информационных изданий, ставших источниками информации в различных отраслях машиностроения. Выделяется роль отраслевых НИИ, активно занимавшихся информационной работой в 30-40-х гг., а также в послевоенное время. Отмечается, что с 1954 г. и по настоящее время ведущая роль в обеспечении различных отраслей машиностроения научно-технической информацией в России принадлежит ВИНИТИ РАН.

Ключевые слова: *реферативный журнал, машиностроение, история реферирования, библиографическая информация*

Первым источником информации об отечественных и зарубежных технических достижениях можно считать журнал «Вестник инженеров», который издавался в 1915–1920 гг. в России – преимущественно аграрной стране со слабо развитым машиностроением. В 1922 г. выпуск журнала был возобновлен. В нем, кроме оригинальных и обзорных статей, публиковалась выборочная библиографическая информация о наиболее интересных статьях из иностранных журналов.

В первые годы Советской власти с началом развития промышленности в стране появился целый ряд изданий – источников информации в области машиностроения. Среди них – журнал «Вестник металлопромышленности», в котором был раздел «Библиография», реферативный журнал «Сообщения о научно-технических работах в Республике», «Реферативный указатель технической литературы», «Новости технической литературы», «Журнальная летопись», ежегодник «Научная литература СССР». В годы первой пятилетки, кроме перечисленных изданий, начинают выходить реферативно-библиографические обозрения по отдельным отраслям машиностроения. В 1927 г. библиотекой Центрального Совета Осоавиахима было начато издание ежемесячного справочно-библиографического обозрения иностранной и советской периодической литературы «Хроника воздушного дела». В 1929 г. Судопроект приступил к выпуску «Бюллетеня», в котором публиковались библиографические описания отечественных и иностранных статей, посвященных различным аспектам судостроения. В 1931 г. Всесоюзное авиационное объединение выпускает «Бюллетень ВАО», в котором освещались различные вопросы работы как отечественной, так и зарубежной авиационной промышленности.

Тогда же в составе ВСНХ был создан специальный сектор технической пропаганды и организован Центральный институт технико-экономической информации, который 1 февраля 1932 г. выпустил первый номер еженедельного бюллетеня технической информации «Новости техники», содержавший разделы: Общее машиностроение; Паровозо-вагоностроение; Автотракторное дело; Судостроение; Авиастроение; Сельскохозяйственное машиностроение; Текстильное машиностроение; Станки и инструмент; Приборы, приспособления, механизмы; Промтранспорт.

В том же 1932 г. Глававиапром вместо «Бюллетеня ВАО» и «Хроник воздушного дела» начинает выпуск реферативно-библиографического журнала «За овладение авиатехникой», который содержал 3 раздела: в первом помещались расширенные рефераты, во втором – международная хроника и в третьем – краткие аннотации по материалам мировой научно-технической периодики. В журнале реферировалось более 40 иностранных первоисточников.

Также в 1932 г. Всесоюзное автотракторное объединение (ВАТО) начало выпуск ежемесячного журнала «За овладение техникой». В 1932-1933 гг. вышло 13 номеров журнала. Тогда же Сектором технической пропаганды Всесоюзного объединения оптико-механической промышленности (ВООМП) стал издаваться «Библиографический указатель иностранной литературы по оптико-механической промышленности», который с 1935 г. стал выходить в качестве приложения к журналу «Оптико-механическая промышленность», где публиковалась информация из 37 иностранных журналов.

В 1933 г. Бюро иностранной технической информации клуба работников народного хозяйства им. Ф. Э. Дзержинского начало подготовку сборника рефератов и аннотаций статей по машиностроению и

металлообработке из примерно 40 первоисточников. С мая 1933 г. до конца 1934 г. вышло семь сборников без строгой периодичности, а с 1935 г. этот сборник стал выходить как журнал рефератов и аннотаций с периодичностью 6 выпусков в год. В 1933 г. появился «Сборник аннотаций по вопросам металлообработки и машиностроения», содержавший сведения как из советских, так и из иностранных журналов.

Информация о достижениях в области горного, нефтяного и насосно-компрессорного машиностроения содержалась в «Библиографическом указателе», который в 1935 г. издавала Центральная конструкторская контора «Гормашпроект». В этот же период выходили «Вестник металлопромышленности» и «Вестник инженеров и техников», в которых публиковались аннотации и рефераты статей из иностранных и некоторых советских журналов.

С 1936 г. научно–технической информацией стала заниматься Государственная научная библиотека Наркомтяжпрома (ГНБ). В 1936–1953 гг. она выпускала библиографический ежемесячник «Новости технической литературы», в котором примерно 15% публикаций сопровождались аннотациями. Ежемесячник выходил шестью сериями**,** крупнейшей из которых была серия «Машиностроение». За первые 10 лет существования ежемесячника было выпущено по 100 номеров каждой из шести отраслевых серий, в которых было помещено около 532 000 сообщений по советской и иностранной технической литературе, причем наибольшее количество из них (около 113 000) – в выпуске «Машиностроение»**.**

В эти годы информационной работой активно занимались ведущие отраслевые НИИ. Например, Центральный институт авиационного моторостроения им. П.И. Баранова (ЦИАМ) в 1938–1949 гг. издавал «Обзорный бюллетень зарубежного авиамоторостроения»; Всероссийский научно-исследовательский институт авиационных материалов (ВИАМ) в 1937–1941 гг. – «Информационно библиографический бюллетень но авиационным материалам» и в 1940–1943 гг. – обзорно-информационный бюллетень иностранной периодики «Авиационные материалы»; Центральный аэрогидродинамический институт им. профессора Н.Е. Жуковского (ЦАГИ) в 1940–1941 гг. выпустил 22 номера «Рефератов статей» из иностранных журналов; в 1940 г. в НИИ машиностроения для текстильной и легкой промышленности был начат выпуск информационного сборника «Новости текстильного и легкого машиностроения».

В годы Великой Отечественной войны издавались «Книжная летопись», «Летопись журнальных статей», журнал «Станки и инструмент» (выделившийся из «Вестника металлопромышленности»), библиографический ежемесячник «Новости технической литературы», а также научно- технический журнал «Вестник машиностроения».

Характерной особенностью послевоенного периода (1946–1954 гг.) стало значительное увеличение количества отраслевых библиографических и реферативных изданий машиностроительного профиля. Самые значительные из них – библиографический справочник «Строительное и дорожное машиностроение» (ВНИИСтройдормаш); «Сборник по обмену опытом заводов строительного и дорожного машиностроения», «Список аннотаций статей по машиностроению из советских и иностранных журналов» (Ленинградский Дом техники машиностроения); «Аннотации статей из иностранных журналов. Технология машиностроения и организация производства» (Оргтрансмаш); «Переводы и рефераты» (Министерство станкостроения СССР); «Рефераты. Технико-информационный бюллетень» (Министерство машиностроения СССР); «Техническая информация» (НАМИ); сборники переводов и обзоров «Современная техника», «Вопросы ракетной техники», «Прикладная механика и машиностроение» (Издательство иностранной литературы).

Со второй половины XX в. в условиях бурного развития науки и техники возросла роль научно-технической информации. Обобщение и анализ накопленных знаний, выбор нужной информации стали неотъемлемыми этапами работ при создании любого инженерного сооружения, машины, аппарата.

Необходимость использования в народном хозяйстве последних достижений науки и техники привела к принятию в нашей стране ряда мер, направленных на усиление роли информационных органов и улучшение их организации.

С 1954 г. централизованную обработку мировой научно–технической литературы осуществлял ВИНИТИ АН СССР. Подготовку информационных изданий по машиностроению Институт начал в 1955 г**.** Первый сводный том РЖ «Машиностроение» вышел в 1956 г. В него вошло около 1000 публикаций как по общим вопросам, так и по отдельным отраслям машиностроения. С 1958 г. кроме сводного тома стали издаваться пять отдельных выпусков. В 1960 г. количество выпусков, входивших в сводный том «Машиностроение», достигло 16, а в 1961 г. – 20. Объем сводного тома при этом увеличился с 40 до 65 авторских листов. Журнал стал не только **большим** по объему, но и **разносторонним** по содержанию.

Начало издания новых специализированных реферативных журналов по машиностроению стало откликом на приоритетные направления развития научно-технического прогресса. Примером этого служит начало издания таких выпусков РЖ, как «Ядерные реакторы» (1958 г.), «Ракетостроение и космическая техника» (1961 г.), «Авиационные и ракетные двигатели» (1961 г.), «Промышленные роботы и манипуляторы» (1982 г.) и т. д.

С 1968 г. ВИНИТИ начал издавать с периодичностью 1 раз в месяц тридцать выпусков РЖ, тематика которых охватила все важнейшие отрасли машиностроения. Общий годовой объем этих выпусков составил 3700 авторских листов и включил 140000 публикаций.

Кроме РЖ, с 1981 г. в ВИНИТИ генерируется база отечественных и зарубежных публикаций по машиностроению. Документы БД содержат библиографические описания, ключевые слова, рубрики и рефераты первоисточников. На основе БД ВИНИТИ по машиностроению по заявкам пользователей подготавливаются любые наборы тематических фрагментов баз данных или их разделов в поисковой системе «Сокол», а также результаты поиска по запросам, выполненным в режиме on-line.

С 1964 г. по 1990 г. отделом научной информации по машиностроению ВИНИТИ издавались серии ежегодников «Итоги науки и техники». За этот период было подготовлено 30 томов серии «Технология машиностроения», 12 томов серии «Авиастроение», 13 томов серии «Судостроение», 12 томов серии «Ракетостроение», 9 томов серии «Машиностроительные материалы, конструкции и расчет деталей машин. Гидропривод», 10 томов серии «Легкая промышленность». Этот вид изданий пользовался огромной популярностью у отечественных и зарубежных специалистов. В настоящее время стоит задача возрождения «Итогов науки и техники» по машиностроению.

Несмотря на сложности последних двух десятилетий, связанные с сокращением входного потока научно-технической литературы и снижением числа потребителей информации, номенклатура издаваемых реферативных журналов по машиностроению практически полностью сохранена. В настоящее время годовой объем выпусков РЖ ВИНИТИ РАН по машиностроению составляет 1006 учетно-издательских листов и включает 43000 документов.

Кроме Реферативного журнала в печатной форме, ВИНИТИ РАН выпускает электронные версии РЖ по машиностроению, которые полностью соответствуют печатным выпускам. Электронный журнал представляет собой информационную систему, позволяющую пользователю на персональном компьютере просматривать отдельные номера Реферативного журнала. По наполнению и порядку расположения разделов и данных каждый номер ЭлРЖ полностью повторяет соответствующий номер РЖ в печатной форме и снабжен общепринятым для информационных изданий механизмом доступа к описаниям документов.

ВИНИТИ РАН осуществляет также подготовку информационно-аналитических обзоров по различным проблемам машиностроения на договорной основе силами ведущих сотрудников ВИНИТИ – специалистов в определенных областях науки, технологий и техники. Основой для обзоров служит фонд отечественной и зарубежной научно-технической литературы, поступающей в Институт.

Материал поступил в редакцию 29.08.17.

Сведения об авторе

ГРЕЧИКОВ Михаил Игоревич – кандидат технических наук, заведующий Отделением ВИНИТИ РАН, Москва
e-mail:mach@viniti.ru