ФЕДЕРАЛЬНОЕ АГЕНТСТВО НАУЧНЫХ ОРГАНИЗАЦИЙ

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

НАУЧНО • ТЕХНИЧЕСКАЯ
ИНФОРМАЦИЯ

Серия 2. ИНФОРМАЦИОННЫЕ ПРОЦЕССЫ И СИСТЕМЫ

ЕЖЕМЕСЯЧНЫЙ НАУЧНО-ТЕХНИЧЕСКИЙ СБОРНИК

Издается с 1961 г. **№ 7** **Москва 2016**

ОБЩИЙ РАЗДЕЛ

УДК 001.4 : 002.1 – 027.21

А.В. Соколов

Когнитивный подход к документу и документосфере

Рассмотрены проблемы использования когнитивного подхода для раскрытия сущности документа и области существования документов. Обоснована потребность в понятии «документальная сфера» и предложены различные интерпретации этого понятия. Статья написана на стыке информатики, документологии и библиотековедения.

Ключевые слова: *документ, документология, информация, когнитивный подход, понятие, сфера документов, теория*

ВВЕДЕНИЕ

Когнитивный (от латинского «cognitio» – знание, познание) *подход* представляет собой рассмотрение человека и человеческого общества в качестве субъектов, создающих, обрабатывающих, хранящих, передающих и использующих знание. Зарождение когнитивного подхода датируется 60-ми годами прошлого века, когда в США возникает междисциплинарное сообщество, получившее название «*когнитивная наука*», или *«когнитология»*, или *«когнитивистика»* и быстро завоевавшее популярность во всем мире [1]. В семействе когнитивных наук представлены: теория искусственного интеллекта – знаковый продукт компьютерной революции, когнитивная психология, когнитивная лингвистика, нейрофизиология, традиционными предметами изучения которых являются интеллект, мышление, ментальность. Когнитивный подход нашел применение в антропологии, педагогике, исторической науке [2]. В синергетике когнитивные исследования понимаются как

«высокие гуманитарные технологии» и определяются как «способы и алгоритмы достижения целей субъектов, опирающиеся на данные о процессах познания, обучения, коммуникации, обработки информации человеком и животными, на представления нейронауки, на теорию самоорганизации, компьютерные информационные технологии, математическое моделирование элементов сознания» [3, с. 193-194]. При этом утверждается, что «когнитивные технологии в перспективе могут привести к созданию новой мегаотрасли, сравнимой с компьютерной индустрией» [3, с. 212]. Неслучайно в «Новой философской энциклопедии» появилась статья, посвященная «когнитивной науке» (М.: Мысль, 2010. Т. 2, с. 264−265), а Р.С. Гиляревский в свой энциклопедический словарь, посвященный информационной сфере, включил дефиницию «*Когнитивистика* – научная дисциплина, изучающая процессы познания в рамках искусственного интеллекта» [4, с. 126].

Казалось бы, когнитивистика и когнитивные технологии должны были привлечь пристальное внимание профессиональных организаторов и хранителей социально ценного знания, обеспечивающих доступ к культурному наследию нынешних и последующих поколений. Я имею в виду сотрудников информационных служб, издательств, библиотек, архивов, музеев, книжных магазинов, для которых документированное социальное знание является предметом труда, а когнитивные процессы вроде читательской деятельности – повседневным занятием. Имеются, конечно, отдельные достижения, о которых забывать никак нельзя. Так, *библиопсихологию* Н.А. Рубакина (1862–1946) можно назвать одной из первых когнитивных дисциплин. В 90-е годы XX века В.А. Фокеев начал разработку *когнитографической концепции библиографии*, исходя из убеждения, что сущность библиографии заключается не в библиографической информации, а в *библиографическом знании*. Он посвятил определению природы и специфики библиографического знания серию научных статей, фундаментальную монографию [5] и учебное пособие, где утверждал, что библиографическое знание по природе своей «результат познавательной деятельности сознания, отражающий объективную реальность в виде знаковой системы, содействующей освоению книжной культуры человеком и обществом в целом» [6, с. 169].

Одновременно с библиографоведом В.А. Фокеевым в Краснодарской академии культуры библиотековед А.И. Остапов выступил с концепцией *библиотечной когнитологии*, в которой с научно-когнитивистских позиций трактовал информационную, коммуникационную и библиотечную деятельность, а также перспективы их компьютеризации [7]. Однако ни в библиографической, ни в библиотечной науке не сложились научные школы когнитологов. Только в наши дни наметилось долгожданное оживление. *Русская школьная библиотечная ассоциация* во главе с Т.Д. Жуковой провозгласила: школьная библиотека – когнитивный ресурс развития образования, её призвание – выполнять когнитивную миссию [8], а также опубликовала сборник когнитивных библиотечных технологий [9]. Очевидно, что когнитивная миссия преемственно связана с просветительной миссией русской интеллигенции, но традиции просветительства не вписываются в библиотечный маркетинг, они чужды архивным учреждениям, хотя и просматриваются в образовательных моделях музеев, культивирующих «музейную педагогику». Неудивительно, что в стандартных учебниках по направлениям «Документоведение и архивоведение» [10, 18], «Музейное дело и охрана памятников» [11], «Библиотечно-информационная деятельность» [12] никаких упоминаний о когнитивистике не содержится. Игнорирование когнитивного подхода в «науках о документах», изучающих различные виды документальной информации, свидетельствует об их старомодности, с которой нельзя примириться. Поэтому в настоящей статье, имеющей постановочный характер, сосредоточим внимание на двух нетрадиционных документоведческих задачах: во-первых, выявление когнитивной сущности документа; во-вторых, осмысление документосферы как объекта когнитивного подхода.

1. КОГНИТИВНАЯ СУЩНОСТЬ ДОКУМЕНТА

Сущность документа – не новая проблема в документоведении. Известны многочисленные дефиниции (определения), претендующие на раскрытие *сущности*, т. е. истинной природы и глубинного содержания документных *явлений*. Чаще всего речь идет об информационной, либо коммуникационной, либо семиотической, либо управленческой сущности (природе, функции) документов. Когнитивная сущность не упоминается никогда, поскольку не практикуется когнитивный подход. В итоге в современном документоведении не утрачивает актуальности вопрос: «Что есть документ?». Так, директор ВНИИ документоведения и архивного дела М.В. Ларин, спустя почти 50 лет после создания этой головной научной организации, не без горечи признал в одном из своих официальных выступлений: «документоведение как научная дисциплина только еще складывается», «мы находимся лишь в процессе её становления», «предстоит определиться с сущностными характеристиками главных объектов и предметов для изучения – документов и систем документации, выявить свойства и признаки документа, проследить развитие его функций на протяжении времени» [13]. Что же нам известно о сущности документа?

Практикуются узкая и широкая трактовки в зависимости от объема понятия «документ». Узкая трактовка представлена в толковых словарях русского языка, где документ определяется как «деловая бумага, юридический акт, служащий доказательством чего-то, подтверждающий право на что-то» или как «официальное удостоверение личности, пропуск, паспорт» [14]. Она стандартизирована в канцелярском делопроизводстве и архивном деле, о чем свидетельствует ГОСТ ИСО 15489-1-2007: «Документ – зафиксированная на материальном носителе идентифицируемая информация, созданная, полученная и сохраняемая организацией или физическим лицом в качестве доказательства при подтверждении правовых обязательств или деловой деятельности». В недавно изданном учебнике для бакалавров по направ-

лению подготовки «Библиотечно-информационная деятельность» сказано, что содержание понятия «документ» трактуется по-разному в разных областях человеческой деятельности: «для геологов – это образцы пород, привезенные из экспедиции, для археологов – находки из раскопа, для историков – различные письменные свидетельства и предметы быта определенного времени, для делопроизводителя – деловые, внутрифирменные документы» [12, с. 12]. Сущность документа в данных случаях есть достоверность (истинность) его содержания. Она является умопостигаемым, то есть когнитивным, качеством, хотя документоведы о когнитивистике не упоминают.

Широкая трактовка, свойственная общим (универсальным) теориям документа, также не чужда когнитивным процессам, но здесь когнитивность вуалируется информационным и семиотическим подходами. Документалист-кибернетик Г.Г. Воробьев в докторской диссертации «Информационная теория документа», защищенной в 1979 г. на стыке документалистики и информатики, рассматривал документ в системе коммуникации в качестве «посредника между источником (автором) и приемником (реципиентом)» и предложил следующую дефиницию: «под *документом* понимается семантическая информация, выраженная на любом языке и зафиксированная любым способом на любом носителе с целью её обращения в динамической системе, иными словами, все то, что, в принципе, может храниться в архивах, библиотеках, музеях» [15]. Таким образом, всякий документ объявлялся элементом некоторой информационной системы, существующей в глобальной сфере информации (***инфосфере***). Здесь требование достоверности не предъявляется, и, при желании, документом можно считать всякий материальный объект, содержащий в себе следы человеческой мысли или эволюции животного мира.

В 90-е годы информационный подход завоевал всеобщее признание, и нельзя не согласиться с утверждением авторитетного библиотековеда-документолога Юрия Николаевича Столярова: «При всем многообразии подходов к пониманию сущности «документа» полное единство взглядов всё же существует в главном: что документ представляет собой содержащуюся на материальной основе, выраженную в знаковой форме *информацию*» (курсив автора) [16, с. 76]. Далее выясняется, что документ суть «разновидность информации», «частный случай информации» [16, с. 84]. Какой информации? Ю.Н. Столяров, надо полагать, имеет в виду семантическую информацию, представляющую собой «амбивалентный феномен, выражающий духовные смыслы в коммуникабельной знаковой форме», но, к сожалению, не фиксирует это обстоятельство в своей итоговой дефиниции: «*Документ* – это информация на материальном носителе, зафиксированная искусственным способом в знаковой форме» [16, с. 125]. Почему так важно указание на семантико-информационную природу документа? Потому что существуют разные типы информации.

В инфосфере циркулирует *машинная информация* в виде способа передачи машиночитаемых кодов посредством коммуникабельных сигналов, развивается *математическая теория информации,* оперирующая абстрактными формулами, а в других отраслях знания и философии информационный подход реализуется посредством *метафор* – «запомненный выбор одного варианта из нескольких возможных и равноправных», «мера разнообразия», «мера сложности структур», «мера организации», «средство создания порядка из беспорядка», «некий алгоритм, совокупность приемов, правил или сведений», «сущность, сохраняющаяся при вычислимом изоморфизме» и т.п. В этих типах информации когнитивная, т. е. познавательная, направленность не акцентируется. С познанием непосредственно связана только *семантическая информация*, представляющая собой способ выражения духовных смыслов (знаний, умений, эмоций, волевых стимулов, фантазий) человекочитаемыми коммуникабельными знаками. Следовательно, именно семантическая информация суть инструмент *когнитивного* подхода, и наоборот: когнитивный подход возможен только по отношению к *семантическим* информационным процессам. Это принципиально важный вывод, свидетельствующий о том, что есть область, где информационный подход и когнитивный подход к документу в сущности *сливаются* друг с другом. Эта область выявляется благодаря семиотическому подходу, который взяли на вооружение филологи-литературоведы.

Осмысливая условия существования языка и механизм смысловой коммуникации, Ю.М. Лотман ввел понятие ***семиосферы*** как присущего данной культуре семиотического пространства. По его словам, «вне семиосферы нет ни коммуникации, ни языка», «семиосфера – и результат, и условие развития культуры», где происходят «столкновения текстов» [17, с. 250-251]. Если филологическое понятие «текст» отождествить с понятием «документ», то семиосфера предстанет в качестве документосферы, и фундаментальные культурологические выводы, которые сделал Ю.М. Лотман, можно распространить, разумеется, с соответствующими оговорками, на документную коммуникацию. В частности, большой интерес для теории документоведения представляют составленные лидерами Тартуско-Московской научной школы «Тезисы к семиотическому изучению культур», где речь идет о культуре как иерархии семиотических систем, о тексте как целостном знаке и как последовательности знаков, проблеме понимания текста, места в культуре текстов различной давности и др. [17, с. 504-524]. Напрашивается вывод: если, согласно бесспорному утверждению Ю.М. Лотмана, семиосфера есть «результат и условие развитие культуры», то и сферу документов также правомерно рассматривать как «результат и условие развития культуры». Знаки породили документы, поэтому изменения семиотического пространства, например, изобретение письменности или электронной коммуникации, непременно ведут к существенным *когнитивным* сдвигам в менталитете общества.

Учитывая сказанное, я готов согласиться с дефиницией документа, предложенной томским документоведом Н.С. Ларьковым: «***Документ*** *– включенная в социальную коммуникацию семантическая структурированная информация, искусственно закрепленная*

на материальном носителе в стабильной знаковой форме» [18, с. 42]. В этом определении явно выражены такие сущностные аспекты документа, как информационный, семиотический, вещественный (материальная стабильность), коммуникационный (коммуникационное сообщение), а также неявно присутствует когнитивная сущность в виде семантического содержания и знаковой формы. Если же когнитивную сущность документа сделать явной, то получится следующее определение: «*Документ* – включенный в межличностную коммуникацию духовный смысл (продукт духовной деятельности), искусственно закрепленный на материальном носителе в стабильной знаковой форме». В обеих дефинициях неизменно сохраняются материальная основа (вещественный аспект), коммуникационная природа (включенность в коммуникацию в качестве сообщения) и стабильная знаковая форма (семиотический аспект). Учитывается, что семантическая информация и духовный смысл – сущности принципиально разные. Информация – объективна, она включена в социальную коммуникацию, а духовный смысл – субъективен, поскольку включен в персональный коммуникационный процесс. Известен афоризм: у книги столько содержаний, сколько читателей. Этот афоризм справедлив с точки зрения когнитивного подхода к индивидуальной читательской деятельности, но неприемлем при информационном подходе к книжной коммуникации.

Когнитивный личностный процесс связан с *психофизической проблемой*, сформулированной в XVII в. дуалистом Рене Декартом (1596−1650) в виде общефилософской проблемы соотношения мыслящей духовной субстанции и телесной материальной субстанции. Применительно к живому человеку нейробиологи и нейропсихологи сформулировали *психофизиологическую проблему,* изучающую взаимосвязи материального субстрата мозга (сеть нейронов, биохимические и электродинамические процессы) и мира смыслов – знаний, эмоций, желаний, образующих духовную жизнь личности. До сих пор эти взаимосвязи остаются загадочным «черным ящиком», декодировать который пытаются когнитивная психология и когнитивная лингвистика. Когнитивная лингвистика, в частности, «исходит из того, что познавательные механизмы и структуры сознания регулярно выражаются в языке. Поэтому язык признается ценным источником сведений о ментальной «инфраструктуре» человека» [19, с. 22] и методологическим средством разрешения психофизиологической проблемы. Поскольку всякий документ есть овеществленная речь, данный вывод правомерно распространить на сферу документов.

Не приходится сомневаться, что когнитивный подход необходим для понимания психофизиологических процессов чтения документов и для овладения «искусством чтения», о котором почти сто лет тому назад замечательно поведал профессор логики Петроградского университета С.И. Поварнин (1870–1952) [20]. К сожалению, как вынуждена признать Ю.П. Мелентьева, у нас до сих пор не существует «единой научной теории чтения, т.е. целостного понимания его природы» [21, с. 12]. Очевидно, что без обращения к психофизиологическим методам здесь не обойтись, но столь же очевидно, что для разработки читателеведения как комплексной научной дисциплины нужно иметь представление о когнитивной структуре социальной документосферы, являющейся продуктом духовной культуры общества.

2. ДОКУМЕНТОСФЕРА КАК ОБЪЕКТ КОГНИТИВНОГО ПОДХОДА

Если когнитивный подход распространить на человеческое общество, то обнаруживается *социальная ментальность*, которую вдумчивый культуролог А.А. Пелипенко определил как «устойчивую совокупность когнитивных механизмов, закрепляющих в культурной традиции те смысловые конструкции, которые способствуют реализации базовых экзистенциальных интенций для той или иной социальной общности» [22, с. 29-30]. Ментальность, – здесь же поясняет автор, – представляет собой «пространство, в котором психические процессы преобразуются в акты мышления», т. е., с точки зрения когнитивного подхода, ментальность – это проявление духовности данного социума. *Духовность* выражается в способностях к речевой коммуникации, абстрактному мышлению, целенаправленному творчеству, которые именуются *когнитивными способностями*. Когнитивные способности обусловливают развитие духовной культуры общества, воплощением которой является «мир объективного знания», обеспечивающий «устойчивое закрепление в культурной традиции экзистенциальных смыслов». Согласно эпистемологической теории классика современной философии Карла Поппера (1902−1994), *мир объективного знания* – «мир продуктов нашего человеческого сознания», который «включает архитектуру, искусство, литературу, музыку и, что, возможно, наиболее важно, науку и образование» [23, с. 19-20]. Продукты человеческого сознания могут быть овеществлены и приобщены к «миру физических вещей» в форме книг или журналов. Таким образом, созданное когнитивными способностями людей объективное знание овеществляется в виде документальной сферы, проще говоря, «документосферы».

«Документосфера» – неологизм, очень редко встречающийся в научных текстах. Только в одном словаре-справочнике удалось обнаружить дефиницию следующего содержания: «Документосфера – сфера обращения документальной информации. Характеризует состояние, качество документальной памяти человечества. Включает такие коммуникационные явления, как: документальный поток, документальный массив, документальный ресурс, документальный фонд; процессы: документирование, документография, документальное обслуживание и др.» [24]. Упрощая приведенную формулировку, назовем ***документосферой*** ту область социально-культурного пространства, где создаются, передаются, хранятся и используются сообщения смысловой коммуникации, именуемые «документы». Данная формулировка приемлема, несмотря на то, что «сфера» является метафорой, а не логически обоснованным термином. Метафоризация используется в научном дискурсе для обозначения интуитивно понятных, но сложно определяемых яв-

лений. «Документосфера» – это не «документальная система» в смысле «целостность взаимосвязанных элементов», а «область бытия родственных предметов»; кроме того, привлекательно, что она логично вписывается в ряд «геосфера», «биосфера», «инфосфера», «ноосфера» и т.п. Конечно, хотелось бы оперировать однозначным понятием «документ», но науки о документах его не знают. Тем не менее, можно кое-что определенное сказать о документосфере.

Когнитивный подход, ориентированный на смысловое содержание, позволил несколько углубить представление о сущности документа (см. раздел 1), что является стимулом для использования его и при *содержательной типологизации документов*. Скрупулезный анализ российской, украинской, польской документоведческой литературы, выполненный Г.Н. Швецовой-Водкой, показал, что существуют восемь значений понятия «документ», соответствующих восьми типам документов, реально используемых в современной практике [25, 26]. Воспроизведем эту типологию.

Документ I. Любой материальный объект, несущий закрепленную информацию, который можно использовать для изучения какого-либо физического или интеллектуального явления.

Документ II. Результат человеческого труда, предмет материальной культуры человечества, свидетельствующий об уровне развития человеческой цивилизации (орудия труда, оружие, здания, машины и т.д.).

Документ III. Материальный объект, созданный специально для передачи в обществе зафиксированной на нем семантической информации, т. е. материальное воплощение человеческой мысли в виде «документов трех измерений» — модели, макеты, скульптуры, памятники и пр.

Документ IV. Материальный объект, на котором семантическая информация закреплена способом записи (независимо от вида записи), включая произведения письменности и печати, рисунки и гравюры, фотографии и кинофильмы, любой вид записи голоса или изображения.

Документ V. Запись, являющаяся спутником определенной духовной деятельности и отражающая её, например, источники личного происхождения (мемуары, дневники, автобиографии, письма), техническая документация.

Документ VI. Запись, содержащая сведения о юридическом факте, например, материалы конкретно-социологических исследований, статистические данные.

Документ VII. Запись о юридическом факте, имеющая необходимый набор засвидетельствования, типа судебно-следственного дела.

Документ VIII. Запись о юридическом факте, удостоверяющая личность — пропуск, мандат, паспорт.

Приведенная типологическая схема построена эмпирически, её достоинство в том, что она отражает в свете когнитивного подхода реальные духовные потребности общества. Что имеется в виду? Документы I, II, III типа объединяют три общих качества: во-первых, все они служат носителями *научного знания*; во-вторых, все они демонстрируют в качестве знаковой формы не запись, сделанную письменами, а собственный *многомерный образ*, свою телесность; в-третьих, они предназначены для использования в когнитивном процессе *социального познания*. Собственно говоря, документы типа I и II генетически документами не являются, процедура их документирования заключается в содержательном отборе и соответствующем комментировании. Документы типа I представляют Вселенную, включающую «любые материальные объекты», в том числе «ландшафты, представителей животного и растительного мира, предметы быта, архитектурные памятники и т.п. ». Г.Н. Швецова-Водка оговаривает, что «значение Документа I охватывает только такие материальные объекты, которые собраны человеком и представлены для обозрения, ознакомления и изучения» в качестве «экспонатов или музейных предметов» [26, с. 28]. Аналогично объем типа Документ II включает не все множество «предметов материальной культуры человечества», а только некоторые особенно познавательные образцы, отобранные историками-музееведами. Другое дело монументы или панорамы, отнесенные к Документам III и специально созданные в качестве произведений культуры и искусства: здесь каждый документ уникален и единственный в своем роде, хотя имеются копии и репродукции отдельных шедевров. Можно сказать, что назначение документов данных типов – служить средством удовлетворения *когнитивных социальных потребностей*, то есть потребностей общества в познавательной информации.

Для того чтобы когнитивная информация, носителями которой служат документы типа I, II, III, дошла до своего потребителя (П), требуется смысловая (когнитивная) коммуникация по схеме Д → П. Эту коммуникацию осуществляют, главным образом, социальные институты *библиосферы* — институты книгоиздания, журналистики, книжной торговли, библиографии, библиотечного дела. Документы типа IV и V, представляющие собой *записи*, – типичные продукты письменности и полиграфии, конечно, являются носителями *духовных смыслов*, охватывающих не только знания и умения, но и волевые воздействия, эмоции, фантазии, т.е. *все виды смыслов.* Материальные носители этих документов ассоциируются с плоской поверхностью бумаги или её заменителей, а знаковую форму образуют все виды *человекочитаемых знаков и изображений* в кодированном или некодированном виде. Социальное назначение документов типа IV и V – служить средством удовлетворения *всех видов духовных потребностей* в текущей и ретроспективной смысловой информации социальных институтов образования, художественной литературы, науки, политической и социальной жизни.

Триада документов типа VI, VII, VIII – носитель *управленческой*, в том числе юридической, информации, содержание которой суть *волевые воздействия* государственной или местной власти, связанные с потребностями управления обществом, регионом, учреждением, фирмой и пр. Для локализации документов этого типа предусмотрен архивный социальный институт, располагающий обширной сетью

архивов, системой подготовки кадров и научно-исследовательским центром архивоведения в лице ВНИИДАД. Главным назначением этого института является содействие государственной власти и другим властным структурам в их управленческой деятельности. Вместе с тем, архивные фонды являются незаменимым источником буквально во всех исторических исследованиях, поэтому они широко используются для удовлетворения когнитивных потребностей общества.

Таким образом, документосфера является пространством осуществления познавательных, коммуникационных, управленческих процессов, опосредованных документами различного типа. Абстрактную структуру документосферы можно представить в виде следующих ***когнитивных секторов.***

А. *Информационный* (точнее – *семантико-информационный*) *сектор*, обеспечивающий отбор и оперативное распространение новых актуальных смыслов в форме документов или, если нужно, в недокументальной форме. Социальные смыслы создаются вне документосферы, этим заняты такие ***духовно-производственные институты***, как религия, наука, философия, техника, искусство. Документирование (фактически – ввод в документосферу) осуществляют периодические издания, текущая беллетристика, научно-информационные, публицистические, политические, рекламные и тому подобные публикации. Профилю информационного сектора соответствуют сети публичных муниципальных, школьных, детских, научно-технических, военных библиотек. Документирование является предпосылкой и условием последующих когнитивных операций. Для дальнейшего развития семантико-информационного сектора документосферы актуально освоение компьютерных информационных технологий, использование электронной массовой коммуникации и Интернета.

Б. *Познавательный сектор* служит транслятором социально ценных смыслов, оформленных в виде документов, соответствующих принятым технологическим нормам. Передачу документированных знаний и умений из поколения в поколение осуществляет народное образование — начальная, средняя, высшая школа. Традиционным источником социальных смыслов является также редакционно-издательский и полиграфический институт, с которым связаны основные социальные институты духовного производства, в том числе — художественная литература и политическая публицистика. Библиография как поисковая инфраструктура документосферы относится к познавательному сектору, поскольку всякий поиск — процесс выявления неизвестного. Так как «музеи служат делу просвещения и воспитания членов общества» [27], они выполняют в документосфере познавательную функцию.

В. *Сектор социальной памяти* представляет собой документированный символ нации. Он включает национальные универсальные и отраслевые книгохранилища, региональные (краевые, областные) научные библиотеки, органы национальной (государственной) библиографии, сеть архивов и музейную сеть. Строго говоря, историко-культурные памятники в виде шедевров архитектуры, дворцовых комплексов, храмов и т.п. должны войти в этот сектор, ибо они выполняют мемориальную когнитивную функцию.

Г. *Управленческий сектор* обеспечивает законотворческую и политическую деятельность государственной власти, организацию общественной жизни на местах, бюрократическое делопроизводство всех учреждений, фирменный менеджмент и т.п. Именно этот сектор обладает глубочайшими историческими корнями, восходящими к древнеегипетским фараонам и «Законам Хаммурапи» XVI в. до н.э. В качестве управленческих воздействий могут использоваться все виды когнитивных продуктов (от знания до фантазий), воплощенные в различные управленческие документы – от конституции страны до расписания поездов. Разумное управление удовлетворяет потребность общества в *самоорганизации*.

Научная ценность абстрактной структуры документосферы заключается в том, что она раскрывает *когнитивные функции,* выполняемые документосферой, а именно: информационная (переработка семантической информации), познавательная, мемориальная, управленческая (лучше – аксиологическая, или ценностно-ориентационная) функции. Однако мы не можем довольствоваться абстрактной функциональной структурой, потому что реальная документосфера характеризуется не только своим общественным назначением, но и природой образующих её элементов. Элементами документосферы, естественно, являются документы, которые, согласно принятой в разделе 1 дефиниции, есть искусственно созданные коммуникационные сообщения. Коммуникационные сообщения создаются, обрабатываются и передаются социально-коммуникационными институтами. Отсюда проблема институционализации, определяющая технологическое функционирование документосферы.

Институционализация документосферы представляет собой организацию социальных институтов, предназначенных для удовлетворения духовных потребностей общества путем оперирования с документами разных типов. *Социальный институт* мыслится в социологии как «исторически сложившаяся устойчивая форма организации совместной деятельности людей» [28], представляющая собой систему норм и специализированных учреждений, обеспечивающих воспроизводство социальных отношений и профессиональных практик. Согласно учрежденческой интерпретации, социальный институт – это система социально-культурных учреждений, включающая: а) учреждения профильной для данного института профессиональной практики (например, школы, театры, музеи); б) учреждения подготовки кадров (профессиональное образование); в) научно-исследовательские центры (отраслевая наука); г) органы управления, в том числе общественные ассоциации; д) отраслевую коммуникацию (периодические издания, профессиональные семинары и конференции). В современной индустриальной цивилизации ***базовыми*** институтами жизнедеятельности общества считаются: институт *семьи и брака*; *экономические* институты, в том числе частная собственность и рынок; *политические* институты, такие как государство, право, судебная система, полиция; институты *жизнеобеспечения* – здравоохранения, со-

циального обеспечения, культурно-досуговые; *духовно-производственные* институты – институты религии, образования, искусства, литературы, науки, философии.

Помимо базовых, обществу необходимы ***инфраструктурные*** институты, представляющие собой вспомогательные профессиональные системы (совокупность функционально специализированных учреждений, кадров, материально-технических средств), удовлетворяющие потребности базовых институтов. Типичными примерами инфраструктурных институтов могут служить транспортные системы, обслуживающие общественные потребности в транспорте, или топливно-энергетическая инфраструктура. Аналогично документосферу, обеспечивающую документирование, обработку, хранение, распространение, использование социальных смыслов, правомерно интерпретировать как совокупность *инфраструктурных социальных институтов*, исторически сложившуюся для удовлетворения потребностей общества в документной коммуникации. Теперь можно уточнить предложенную ранее исходную дефиницию документосферы следующим образом: ***Документосфера*** *– это область социально-культурного пространства, в которой функционирует система инфраструктурных документо-коммуникационных институтов.*

Документо-коммуникационные институты возникали в процессе эволюции документальных коммуникаций, основные вехи которой, начиная с крещения Руси, хорошо известны. В XX в. сформировались три относительно самостоятельные ***отрасли культуры,*** включающие документно-коммуникационные институты различной целевой направленности, отличающиеся концентрацией практики на документах различного типа, специфической профессионализацией кадров, собственными историческими традициями и перспективами информатизации. Эти отрасли обладают прикладной, практической направленностью, что свойственно инфраструктурным образованиям, поэтому назовем их по имени того конкретного «дела», которое их породило. Тогда получаем следующую *отраслевую конфигурацию* документосферы:

1. *Книжное дело* – система институтов книгоиздания и журналистики, книжной торговли, библиотечного и библиографического институтов, базирующаяся на машинной полиграфии; эту отрасль, учитывая её культурно-историческую значимость, резонно именовать ***библиосферой***.

2. *Архивное дело* – архивный социальный институт, работающий с деловой документацией, отобранной для долговременного хранения.

3. *Музейное дело* – музейный социальный институт, который осуществляет документирование объектов природы и созданных человеком музейных предметов, комплектует музейные фонды, занимается их хранением и изучением, а также выполняет функции образования и воспитания посетителей музея.

Нетрудно удостовериться в том, что архивные, книжные, музейные социальные институты соответствуют учрежденческой интерпретации и образуют разветвленные отрасли культуры, объединяющие десятки тысяч учреждений и более сотни тысяч профессиональных работников. Весьма высок интеллектуальный уровень кадровых ресурсов: все отрасли документосферы имеют вузы, обеспечивающие подготовку бакалавров и магистров, сложились научные школы, насчитывающие десятки кандидатов и докторов наук, на страницах отраслевых научных журналов и продолжающихся трудов конференций из года в год обсуждаются актуальные исторические и текущие социально-культурные проблемы, среди которых неизменно дебатируется вопрос «что есть документ?», но никогда не поднимается вопрос о документосфере и когнитивном подходе к документной коммуникации.

В настоящей статье мы попытались привлечь внимание коллег к научно-методологическим проблемам в области документо-коммуникационных наук, которые требуют межотраслевого, научно-интегрального подхода и преодоления ведомственно-отраслевой замкнутости. В этой связи представляется актуальным вопрос о формировании ***когнитивной документологии,*** *объектом* которой была бы документосфера постиндустриальной цивилизации; *предметом* – когнитивные процессы в постиндустриальной документосфере, а *целью* – сохранение национального культурного наследия в эпоху господства конвергентных нано-био-инфо-когнитивных технологий. Однако обсуждение проблематики когнитивной документологии выходит за пределы настоящей статьи.

СПИСОК ЛИТЕРАТУРЫ

1. Когнитивный подход. Научная монография / отв. ред. В.А. Лекторский. – М. : РООИ «Реабилитация», 2008. – 464 с.
2. Медушевская О.М. Теория и методология когнитивной истории. – М. : РГГУ, 2008. – 358 с.
3. Малинецкий Г.Г. Чтоб сказку сделать былью… Высокие технологии – путь России в будущее. – М. : Книжный дом «ЛИБРОКОМ», 2013.
4. Гиляревский Р.С. Информационная сфера. Краткий энциклопедический словарь. – СПб. : Профессия, 2016.
5. Фокеев В.А. Природа библиографического знания: монография. – М.: РГБ, 1995. – 351 с.
6. Фокеев В.А. Библиография: теоретико-методологические основания: учеб. пособие. – СПб.: Профессия, 2006.
7. Остапов А.И. Введение в библиотечную когнитологию: учебное пособие по спецкурсу. – Краснодар: КГАК, 1994. – 331 с.; Он же. Библиотечная когнитология: монография. – Краснодар: КГАК, 1995. – 330 с.
8. Жукова Т.Д. Когнитивная миссия школьной библиотеки // Информация и научное мировоззрение: сб. статей. – М. : РШБА, 2013. – С. 61–78.
9. Сайкс Дж. А. Школьные библиотеки, дружественные мозгу. Серия. В помощь педагогу-библиотекарю. Вып. 7. – М.: РШБА, 2014. – 152 с.
10. Документоведение: учебник для вузов / под ред. М.В. Ларина. – М. : Академия, 2016. – 320 с.

11. Основы музееведения: учебное пособие / отв. ред. Э.А. Шулепова. – М.: Книжный дом ЛИБРОКОМ, 2010. – 432 с.
12. Гордукалова Г.Ф., Захарчук Т.В., Плешкевич Е.А. Документоведение. Часть 1. Общее документоведение: учебник / науч. ред. Г.В. Михеева.. – СПб.: Профессия, 2013. – 320 с.
13. Ларин М.В. Актуальные проблемы современного документоведения // Вестник РГГУ. – 2014. – № 2. – С. 139.
14. Толково-энциклопедический словарь. – СПб.: Норинт, 2006. – С. 557.
15. Воробьев Г.Г. Информационная теория документа. Автореферат дисс. …. доктора технич. наук. – М.: МГИАИ, 1979. – С. 6.
16. Столяров Ю.Н. Документология: учебное пособие. – Орёл: Горизонт, 2013.
17. Лотман Ю.М. Семиосфера. Культура и взрыв. Внутри мыслящих миров. Статьи. Исследования. Заметки. – СПб. : Искусство-СПБ, 2000.
18. Ларьков Н.С. Документоведение: учебник. – М.: Проспект, 2016. – 416 с.
19. Скребцова Т.Г. Когнитивная лингвистика. Курс лекций. – СПб.: Филологический факультет СПбГУ, 2011.
20. Поварнин С.И. Как читать книги для самообразования // Поварнин С.И. Сочинения. – СПб.: Институт иностранных языков, 2015. – С. 649-698.
21. Мелентьева Ю.П. Общая теория чтения. – М.: Наука, 2015.
22. Пелипенко А.А. Избранные работы по теории культуры. Культура и смысл. – М.: ООО Изд-во «Согласие»; изд-во «Артем», 2014.
23. Поппер К.Р. Знание и психофизическая проблема: В защиту взаимодействия. – М.: Изд-во ЛКИ, – 2008..
24. Ильганаева В.А. Социальные коммуникации (теория, методология, деятельность): словарь-справочник. – Харьков: «Городская типография», 2009. – С. 83.
25. Швецова-Водка Г.Н. Общая теория документа и книги: учебное пособие. – М.: Рыбари; К.: Знання, 2009. – С. 14–37.
26. Швецова-Водка Г.Н. Документ в свете ноокоммуникологии: научно-практическое пособие. – М.: Литера, 2010. – С. 22-167.
37. Философия музея: учеб. пособие / под ред. М.Б. Пиотровского. – М.: Инфра-М, 2013. – С. 5.
28. Глотов М.Б. Социальный институт: определение, структура, классификация // Социологические исследования. – 2003. – № 10. – С. 13-19.

Материал поступил в редакцию 26.04.16.

Сведения об авторе

СОКОЛОВ Аркадий Васильевич – доктор педагогических наук, профессор кафедры информационных и мультимедиа систем Санкт-Петербургского государственного института культуры
e-mail: sokolov1.spb@gmail.com

УДК 004.021.056.55

А.Ю. Щербаков, М.Р. Биктимиров

Системно-аналитический подход к оптимизации алгоритма криптографического преобразования «Кузнечик»

Обсуждается проблема максимальной оптимизации алгоритма криптографического преобразования «Кузнечик». Представлен системно-аналитический подход к оптимизации этого алгоритма.

Ключевые слова: *криптография, алгоритм «Магма», алгоритм «Кузнечик», системный анализ, синтез преобразований*

Национальным стандартом Российской Федерации ГОСТ Р 34.12–2015 [1] определены два базовых алгоритма шифрования – «Кузнечик» и «Магма», которые могут применяться в криптографических методах обработки и защиты информации, в том числе для обеспечения конфиденциальности, аутентичности и целостности информации при ее передаче, обработке и хранении в автоматизированных системах [2]. Настоящая работа, посвященная проблеме максимальной оптимизации нового криптографического алгоритма «Кузнечик» проведена в рамках разработки математических моделей макроэкономической динамики в информационно-платежных системах нового поколения по программе фундаментальных исследований Отделения математических наук РАН «Алгебраические и комбинаторные методы математической кибернетики и информационные системы нового поколения». Смысл работы составляет системно-аналитический анализ алгоритма с переходом к его эквивалентным представлениям в рамках синтеза преобразований в одну системную целостность.

В отличие от ГОСТ 28147-89 (Магма) [3] новый алгоритм представляет собой не сеть Фейстеля, а SP-сеть – преобразование, состоящее из нескольких одинаковых раундов, при этом каждый раунд состоит из нелинейного и линейного преобразований, а также операции «наложения» ключа. В отличие от сети Фейстеля, при использовании SP-сети преобразуется весь входной блок, а не его половина. Такая структура иногда также называется AES-like (похожей на AES), однако, в отличие от последнего, у алгоритма «Кузнечик» есть ряд особенностей:

1) линейное преобразование может быть реализовано с помощью регистра сдвига;

2) ключевая развертка реализована с помощью сети Фейстеля, в которой в качестве функции используется раундовое преобразование исходного алгоритма.

Рассмотрим базовое преобразование алгоритма «Кузнечик» (зашифрование):

```
int funcLSX(unsigned char* a, unsigned char* b,
unsigned char* outdata)
{
    unsigned char temp1[16];
    unsigned char temp2[16];
    funcX(a, b, temp1);
    funcS(temp1, temp2);
    funcL(temp2, outdata).
}
```

При этом преобразование X – побайтное сложение по mod2 ключа и открытого текста, преобразование S – байтовая подстановка, а преобразование L – шестнадцатикратное преобразование при помощи регистра сдвига.

Предельная оптимизация (т.е. оптимизация, при которой невозможно улучшение производительности алгоритма при разумных затратах памяти) состоит в том, чтобы предвычислить одновременно преобразования S и L последовательно для каждого байта вектора длиной 16 байт, при этом каждый байт принимает значение от 0 до 255, а остальные байты вектора равны 0. Таким образом, за счет «байтовости» подстановки и линейной независимости векторов можно «собрать» итоговый вектор тем же аддитивным преобразованием – суммой по mod2.

Поясним это процедурой генерации собственно данных в виде массивов tSL00, … tSL15:

```
for(i=0;i<16;i++)
      {
      printf("unsigned char tSL%02d[256][16]=\n",i);
      printf("{\n");
      for(j=0;j<256;j++).
       }
```

```
        {
          for(m=0;m<16;m++) in[m]=0;
          in[i]=kPi[j];
          funcL(in,out);
          for(m=0;m<16;m++)
printf("0x%02x,",out[m]);
          printf("\n");
         }
        printf("};\n").
```

Для контроля приведем две строки массива tSL00:

```
unsigned char tSL00[256][16]=
{
0xe9,0xfb,0xd5,0x0c,0x7a,0xc0,0x80,0x96,0x19,
0x11,0x87,0x93,0x1b,0xc9,0xae,0xb5,
0x19,0x08,0xe0,0x8c,0xb2,0x17,0x1a,0xce,0x7b,
0x32,0xfc,0xab,0xf8,0xfe,0xf2,0x0a.
}
```

Таким образом, одна итерация алгоритма превращается в последовательность действий:

```
int funcLSX(unsigned char* a, unsigned char* b,
unsigned char* outdata)
{
   unsigned char temp1[16];
   funcX(a, b, temp1);
   funcSL(temp1, outdata);
   return 0.
}
```

Функция funcSL приведена для первого массива и должна быть повторена 16 раз (далее приведем однократную операцию):

```
int funcSL(unsigned char* indata,unsigned char*
outdata)
      {
      int i;
//    for(i=0;i<16;i++) outdata[i]=
tSL00[indata[ 0]][i];
      *(unsigned long *)  outdata   =*(unsigned long *)
tSL00[indata[ 0]];
      *(unsigned long *) (outdata+4 )=*(unsigned
long *)(tSL00[indata[ 0]]+4);
      *(unsigned long *) (outdata+8 )=*(unsigned
long *)(tSL00[indata[ 0]]+8);
      *(unsigned long *) (outdata+12)=*(unsigned
long *)(tSL00[indata[ 0]]+12).
      }
```

Закоментаренная часть не дает существенного увеличения скорости по сравнению с реализацией «в лоб». Однако работа с длинными блоками данных показывает существенный рост скорости.

Таким образом, алгоритм раскладывается как суммирование по mod2 с раундовым ключом, а затем – как суммирование предвыбранных по индексу шифруемого байта 16-и байтных массивов.

Возможно, такой изоморфизм алгоритма может быть использован и для уточненного криптоанализа.

В любом случае, легко видеть, что это максимальная оптимизация алгоритма зашифрования. Дополнительной ценностью алгоритма «Кузнечик» при реализации его в варианте, устойчивом к перехвату ключа по ПЭМИН*, является то, что маска, наложенная на ключ, может быть снята не после суммирования ключа под маской и открытого текста, а после завершения раунда преобразования, что практически позволит полностью избежать появления ключа в памяти компьютера в открытом виде.

СПИСОК ЛИТЕРАТУРЫ

1. ГОСТ Р 34.12–2015. Информационная технология. Криптографическая защита информации. Блочные шифры.
2. Бондаренко А., Маршалко Г., Шишкин В. ГОСТ Р 34.12–2015: чего ожидать от нового стандарта? // Information Security/Информационная безопасность. – 2015. – № 4. – С. 48–50.
3. ГОСТ 28147–89 (Магма). Системы обработки информации. Защита криптографическая. Алгоритм криптографического преобразования.

Материал поступил в редакцию 15.06.16.

Сведения об авторах

ЩЕРБАКОВ Андрей Юрьевич – доктор технических наук, профессор НИУ ВШЭ, Федеральный исследовательский центр «Информатика и управление» РАН, Москва
e-mail: x509@ras.ru

БИКТИМИРОВ Марат Рамильевич – кандидат технических наук, профессор НИУ ВШЭ, ВРИО директора ВИНИТИ РАН, Москва
e-mail: marat@ras.ru

* ПЭМИН – побочные электромагнитные излучения и наводки

ИНФОРМАЦИОННЫЕ СИСТЕМЫ

УДК 351(470) : [001.891 : 002 – 047.44]

П.А. Калачихин

Принципы построения государственной наукометрической системы*

Обсуждается методология создания государственной наукометрической системы в Российской Федерации. Происходит обращение к зарубежному опыту создания наукометрических систем. Определяются цели и задачи создания российской наукометрической системы. Формулируются принципы отбора и предлагается набор наукометрических показателей для включения в отечественную наукометрическую систему.

Ключевые слова: *наукометрические показатели, научно-инновационная деятельность, научно-техническая информация, управление наукой.*

ВВЕДЕНИЕ

Растущая потребность российской экономики в инновационной продукции собственного производства, с одной стороны, и недостаток государственного финансирования, с другой стороны, предопределили необходимость российских ученых сосредоточиться на коммерциализации своих разработок. Мотивы осуществления научно-исследовательской деятельности претерпели изменения. В дополнение к прежнему стремлению получать выдающиеся результаты в научных изысканиях, зарабатывая признание в кругу научного сообщества, большинство ученых стали преследовать коммерческие интересы, что является признаком рационального экономического поведения. Именно поэтому большое распространение получили различные показатели, на основании которых производится оценка деятельности отдельных ученых, научных организаций или научных изданий. В связи с проводимой реформой науки большинство субъектов научной деятельности оказались вовлечены в погоню за высокими значениями показателей эффективности собственной деятельности, потому что на их основании формируются рейтинги конкурсов на получение денежных грантов, всевозможными показателями пронизаны квалификационные требования к соисканию ученых степеней и званий, научных должностей. Интересы ученых были переплетены с количественными показателями, в улучшении которых теперь заинтересован каждый российский ученый.

Прежде чем продолжать изложение, следует привести определения некоторых терминов, часто упоминаемых в дальнейшем.

Наукометрия – это дисциплина, изучающая эволюцию науки через многочисленные измерения и статистическую обработку научной информации.

Наукометрические показатели – индексы публикационной активности авторов или организаций, значимости публикаций в зависимости от научного веса журнала и прочих характеристик научной деятельности. Наукометрические показатели используются для оценки состояния и перспективности научно-исследовательской деятельности авторов и организаций, их сравнения и ранжирования в различных рейтингах.

Наукометрические системы – это библиографические и реферативные базы данных, занимающиеся сбором и обработкой научной информации, являющиеся инструментом для оценки результативности и эффективности деятельности научно-исследовательских организаций, ученых, отслеживания цитируемости научных статей, определения импакт-фактора журнала и других наукометрических параметров.

Сложившееся в отечественной науке положение дел говорит о необходимости создания государственной наукометрической системы (*ГНС*). Глобальной целью создания ГНС является проведение оценок и осуществление прогнозов развития науки, при помощи которых будут отслеживаться состояние и тенденции развития отечественной науки. На основании собранных данных служащие государственных структур в области управления наукой смогут принимать своевременные и адекватные меры по спасению науки.

Актуальность рассмотрения принципов построения ГНС в настоящей статье обусловлена тем, что существующий механизм оценки результатов научной деятельности устарел морально, поэтому имеет смысл заняться созданием ГНС прямо с "чистого

* Работа выполнена в рамках исследования по теме 0003-2015-0008 Госзадания ВИНИТИ РАН

листа". Разработка системы не требует демонтажа существующего механизма ввиду того, что такой механизм попросту не функционирует в результате социальных и экономических потрясений последних нескольких десятилетий. В отличие от значительного количества других публикаций по наукометрии, содержащих полемику, дискуссионные комментарии и соображения, имеющие малую практическую ценность, наше исследование содержит предложения для принятия конкретных действий и реализации конкретных проектов.

Цель настоящего исследования – построение принципов создания государственной наукометрической системы на основании приоритетов развития науки и техники с учетом международного опыта. Обсуждение принципов создания ГНС распадается на ряд отдельных задач, среди которых: формулировка целей создания, выявление круга пользователей и заказчиков, выявление принципов отбора показателей, составление списка таких показателей и задание критериев выбора параметров для их оценки.

НАУКОМЕТРИЧЕСКИЕ СИСТЕМЫ В РОССИИ И ЗА РУБЕЖОМ

На сегодняшний день в отечественной науке не существует единой государственной наукометрической системы, но это вовсе не означает, что полностью отсутствуют инструменты для оценки наукометрических показателей, характеризующих состояние науки в нашей стране. На базе web-платформы создан Российский индекс научного цитирования (РИНЦ), предназначенный для измерения продуктивности научной деятельности авторов публикаций в периодических научных изданиях. Функционирует электронный ресурс "Карта российской науки", содержащий сведения о положении дел в различных областях отечественной науки. Сравнение Российской Федерации с другими развитыми державами, такими как Соединенные Штаты Америки и Китайская Народная Республика, приводит к выводу, что отечественная наукометрическая система находится еще в недостаточно зрелом состоянии. Вопрос о перспективах развития государственной наукометрической системы представляется в этом свете актуальным и стоит на повестке сегодняшнего дня.

Не совсем корректно рассматривать ГНС как совокупность всех существующих наукометрических показателей, методик их оценки и программного обеспечения для научных информационных систем, так как имеет принципиальное значение, кем именно и на основании каких данных рассчитываются наукометрические показатели. Этим могут заниматься совершенно разные организации, такие как фонды развития науки, научно-исследовательские институты или независимые организации по частной инициативе. Говорить об общегосударственной наукометрической системе в таком случае будет не уместно.

РИНЦ в своей базе данных содержит несколько миллионов публикаций, но это всего лишь верхушка информационного айсберга. По грубым оценкам, массив информации, к которому обращается РИНЦ, составляет не более чем 20% от всех научных публикаций российских авторов, в то время как оставшиеся 80% не индексируются и скрыты от прямого доступа. При этом в РИНЦ по понятным причинам совершенно не учитываются публикации иностранных авторов, в том числе написанные на других языках, например, на китайском.

Причинами создания китайской наукометрической системы послужили два условия: наличие исходных данных и большой общественный спрос. Государственная наукометрическая система Китайской Народной Республики построена таким образом, что каждая публикация приписывается единственной научной организации, как заявлено авторами. Если имеется больше, чем одна организация, используется организация, указанная первой. Публикация оценивается только однократно. Публикации классифицируются ручным способом по классификаторам, действующим в рамках китайских стандартов. Для каждой публикации рассматриваются следующие вопросы и присуждаются соответствующие рубрики, в зависимости от полученного ответа: *финансируется ли данная публикация в рамках проекта Национального фонда естествознания Китайской Народной Республики или других крупнейших фондов*? *Имеются ли иностранные авторы?* За временной промежуток длиною в несколько лет кодируется такая информация, как задержка между принятием рукописи и выпуском публикации в печать; возраст, пол, профессиональная должность первого автора и другие [1]. Китайская Народная Республика взяла курс на максимальное расширение присутствия в англоязычных международных журналах. Китайских ученых премируют за публикации в англоязычных журналах [2].

Существует набор показателей для библиометрических и наукометрических оценок результативности деятельности китайских ученых. В статистическом сегменте китайских наукометрических показателей содержится статистика по распределению китайских статей по журналам, регионам и странам. Здесь также приводятся данные о рейтинге зон журналов, содержащих статьи китайских ученых. В этом же разделе содержится распределение публикаций статей и цитат по разным тематикам. Это подмножество показателей является основой для всех остальных наукометрических показателей. Институциональные показатели количества статей и цитирований извлекаются для всех видов китайских научных организаций. Китайская наукометрическая система содержит набор показателей, в том числе для государственных и частных лабораторий. Количество статей в таких лабораториях, число цитирований, число авторов, предмет распределения статей и количество статей, поддерживаемых китайской наукометрической системой, подсчитываются и сводятся в таблицу. Количество основных институтов в каждом регионе Китайской Народной Республики представлены на основе данных из индикаторов региональных наукометрических показателей. Эти данные используются для анализа дисциплинарного преимущества и недостатков научной деятельности в каждом регионе Китая. Показатели деятельности научно-технических фондов играют важную роль в финансировании фундаментальных исследований. Распределение по полу,

возрасту, ученой степени китайских ученых на основании данных, полученных из первоисточников, собираются и используются для анализа и оценки социологических особенностей сообщества китайских ученых и образуют группу авторских наукометрических показателей [3].

Ситуация с наукометрическими показателями в зарубежной и отечественной науке кардинально различается. Отрицательные черты свойственны зарубежной науке, отечественная наука страдает от более серьезных проблем, но имеющих другой тип, поэтому роль наукометрических показателей в сложившихся условиях позитивна. При этом нельзя отрицать, что РИНЦ далек от совершенства [4].

ЦЕЛИ И ЗАДАЧИ СОЗДАНИЯ ГОСУДАРСТВЕНОЙ НАУКОМЕТРИЧЕСКОЙ СИСТЕМЫ

В соответствии со Стратегией инновационного развития Российской Федерации на период до 2020 г. [5] в среднесрочной перспективе должны быть достигнуты следующие цели по развитию отечественной науки:

- увеличение количества цитирований в расчете на одну публикацию российских исследователей в научных журналах, индексируемых в базе данных "Сеть науки" (Web of Science), до четырех ссылок к 2020 г. [5];
- доля России в общемировом количестве публикаций в научных журналах, индексируемых в базе данных "Сеть науки" (WEB of Science), к 2020 г. должна составить 3% [5];
- увеличение количества патентов, ежегодно регистрируемых российскими физическими и юридическими лицами в патентных ведомствах Европейского союза, Соединенных Штатов Америки и Японии, до 2,5-3 тыс. патентов к 2020 г. [5].

Развитие отечественной науки должно выразиться в максимизации следующих наукометрических показателей в течение нескольких ближайших лет:

- число публикаций российских авторов в научных журналах, индексируемых в базе данных Scopus, в расчете на 100 исследователей [6];
- коэффициент изобретательской активности (число отечественных патентных заявок на изобретения, поданных в России в расчете на 10 тыс. человек населения) [6];
- удельный вес публикаций в соавторстве с зарубежными учеными в общем числе публикаций российских авторов в научных журналах, индексируемых в базе данных Scopus [6].

Создание ГНС не в состоянии обеспечить достижение всех обозначенных целей и максимизировать все указанные индикаторы, поскольку, в конечном счете, все зависит от продуктивности научной деятельности отечественных ученых и организаций. ГНС должна внести определенный вклад в развитие отечественной науки, опосредованно влияя на целевые показатели.

Создание государственной наукометрической системы преследует решение задач, способных глобально повлиять на эффективность отечественной науки, среди которых следует выделить:

♦ увеличение в целом продуктивности результатов функционирования отечественной науки в виде повышения наукометрических показателей;

♦ повышение степени информированности авторов и организаций о своем уровне и признании в национальном и международном научном сообществе;

♦ предоставление удобного доступа на определенных условиях всем желающим и заинтересованным лицам к информации из государственных наукометрических баз данных;

♦ обеспечение контроля выполнения квалификационных требований присвоения ученых степеней, званий и научных должностей в соответствии с принятыми законодательными постановлениями;

♦ повышение степени адекватности выбора научных изданий для осуществления публикаций отечественными учеными;

♦ определение приоритетов, проведение конкурсов, присуждение наград на основании оценки достигнутых результатов отечественными учеными и имеющими отношение к науке лицами методом ранжирования по наукометрическим показателям;

♦ достижение и поддержание высокого уровня качества представляемых материалов научных исследований в публикациях;

♦ отсеивание низкокачественных и лженаучных публикаций, содержащих ошибочные результаты или препятствующих развитию науки;

♦ поощрение наиболее передовых в научном плане организаций и ученых в качестве образцовых примеров и мотивация достижения высоких результатов заданного уровня остальными учеными;

♦ формирование ряда показателей по государственной статистике на основании материалов регулярных отчетов о состоянии наукометрических показателей в различных информационных аспектах;

♦ мониторинг состояния ключевых областей науки, в том числе отслеживание активности в отношении областей науки, имеющих особый статус или особое значение.

Выполнение всех этих задач станет возможным благодаря тому, что отечественные ученые получат возможность пользоваться сервисами ГНС. Мотивацию к использованию системы возможно усилить, учитывая требования, законы, инструкции и другие формальные обязательства, тем самым сделав использование сервисов ГНС необходимым условием для проведения научной деятельности в Российской Федерации.

Заказчиками создания ГНС являются государственные ведомства Министерства образования и науки РФ и Федерального агентства по науке и образованию (ФАНО). ГНС предназначается для всех лиц, вовлеченных в науку (аспирантов, докторантов, преподавателей, научных сотрудников); для научных организаций (вузов, научно-исследовательских институтов, фондов поддержки науки); для государственных служащих в сфере науки и образования; для печатных изданий (журналов, издательств) и др.

Существуют полярные точки зрения о пользе оценки наукометрических показателей для развития

науки и на то, какие наукометрические показатели необходимо оценивать и какие математические модели для этого использовать. Создание ГНС не должно нарушать интересы всех ученых, даже если они придерживаются различных взглядов на наукометрию. Создавая ГНС, необходимо найти компромисс между противниками и сторонниками наукометрии, среди приверженцев библиометрических и экспертных наукометрических показателей. Построение ГНС необходимо осуществлять таким образом, чтобы не навязывать лишних обязанностей отечественным ученым, оставляя им свободу действий.

ПРИНЦИПЫ ОТБОРА ПОКАЗАТЕЛЕЙ ДЛЯ ВКЛЮЧЕНИЯ В ГОСУДАРСТВЕННУЮ НАУКОМЕТРИЧЕСКУЮ СИСТЕМУ

Состав наукометрических показателей для включения в ГНС является предметом для дискуссии, более того, существуют разные точки зрения на выбор подхода к оценке наукометрических показателей в ГНС. Например, имеется инициатива использовать экспертные оценки для формирования наукометрических показателей. Вместе с экспертными оценками хорошо было бы использовать нечеткую логику, но в таком случае процедура оценки наукометрических показателей становится слишком трудоемкой и запутанной, поэтому от этой идеи приходится отказаться. По этой же причине не стоит разрабатывать для наукометрических показателей ГНС сложные рейтинговые модели. Совсем другой вопрос – насколько много наукометрических показателей должно быть включено в ГНС, по которому также имеются различные мнения.

Для отбора показателей ГНС, взвесив все за и против, предлагается использовать следующие принципы:

- методики оценки показателей, записанные в виде формул, должны быть простыми и наглядными;
- значения показателей должны быть интерпретируемыми;
- показатели не должны повторяться, необходимо исключить их дублирование;
- значения показателей должны быть оптимизируемы, т.е. стремление показателя к максимуму (или к минимуму) должно говорить о качественных изменениях в лучшую или худшую сторону;
- отбор показателей должен быть подчинен целям и задачам создания ГНС;
- оценка показателей должна давать полезные для науки результаты;
- необходимо укладываться в системные ограничения по объему системных ресурсов, необходимых для их оценки и хранения в базе данных;
- показатели должны обладать новизной, при этом следует избегать использования показателей, защищенных авторским правом, являющихся чужими разработками;
- показатели должны быть приемлемы для ввода в базу данных;
- показатели должны оцениваться периодически, требуя умеренной частоты перерасчетов;
- показатели должны быть способны рассчитываться автоматически или полуавтоматически с минимальным участием человека;
- показатели должны быть удобно извлекаемы из исходных данных;
- показатели должны коррелировать с макропоказателями, такими, как суммарное количество публикаций по стране.

Научное сообщество не имеет согласованных критериев оценки качества научной деятельности. Применительно к конкретному ученому оценка качества научной деятельности реализуется в виде индивидуальных решений, совместных голосований и других аналогичных форм. Все используемые критерии оценки качества научной деятельности обладают субъективным характером. Методология объективной оценки качества научной деятельности должна имитировать процедуру выявления победителей в конкурсах, осуществляемую с ориентацией на имеющиеся достижения, выраженные количественными показателями [7].

ПОДХОДЫ К ОЦЕНКЕ НАУКОМЕТРИЧЕСКИХ ПОКАЗАТЕЛЕЙ

Данные зарубежных наукометрических служб и систем, связанных с цитированиями, являются количественными и качественно интерпретируемыми показателями, с помощью которых на сегодняшний день возможно объективно оценивать результаты и эффективность деятельности отечественных ученых [8].

Подходов к оценке наукометрических показателей существует всего три: подход, основанный на библиометрических показателях; формальный подход и подход, основанный на экспертных методах.

Библиометрические показатели применяются только для тех областей научной или научно-технической деятельности, результаты которых описываются в научных статьях или иных научных публикациях, т.е. преимущественно для фундаментальных исследований и в какой-то мере для прикладных научных исследований, но не для разработок. Для этих областей деятельности более адекватным измерителем будут патенты или какие-либо иные практические результаты, которые по значимости для оценки сравнимы с научными статьями или монографиями, но не поддаются стандартным методам библиометрического анализа.

Следующие количественные библиометрические показатели доступны при проведении анализа результативности НИР с использованием баз данных цитирований:

- количество публикаций конкретного автора, организации, страны и их распределение по журналам и областям знаний;
- распределение данных публикаций по годам издания и по соавторам, характеризующее научные связи ученого или организации;
- количество цитирований всех статей и каждой статьи отдельно с возможностью просмотра перечня цитирующих статей и их дальнейшего анализа.

Библиометрические показатели, являясь по своей природе статистическими, хорошо работают на

больших массивах публикаций, что позволяет достаточно адекватно сравнивать научную деятельность, например, по странам. Использование библиометрических показателей на более детальном уровне анализа научной деятельности оказывается полезным только в сочетании с другими показателями результативности научной деятельности. Возможность и способы использования библиометрических показателей для оценки научной деятельности в значительной степени зависит от целей, с которыми проводится оценка, и должны сочетаться с другими показателями и экспертной оценкой [9].

Существует большое количество библиометрических показателей, рассчитываемых формальным способом. Ученый обладает высокими формальными показателями, если в числе используемых библиометрических индексов есть такие показатели, которые персонально для него значительно выше среднего уровня. Низкая оценка формальных показателей означает, что все основные показатели деятельности ученого существенно ниже среднего уровня [10].

Оценка наукометрических показателей на основании библиометрических данных приводит к ряду методологических ошибок. Первый тип методологических ошибок выражается в неоправданном упоре на числе публикаций и цитирований в научных журналах при оценивании эффективности научной деятельности ученых и организаций. Для разработки адекватных методов такой оценки необходимо проследить развитие научных результатов.

Методологически ошибочными являются попытки оценивать научную продуктивность коллективов и отдельных исследований только на основе публикаций в журналах. Методологические ошибки при упоре на индексы цитирования приводят к неправильным управленческим решениям. Не получают адекватной оценки новые научные направления. Вне оценивания оказываются наиболее ценные результаты. Оценка по импакт-фактору объективно задерживает подготовку печатных изданий. Управление наукой на основе использования в государственной наукометрической системе большого количества публикаций и индексов цитирования объективно замедляет развитие науки и переход полученных научных результатов в область практического применения.

Ссылки на научные публикации производятся не только при составлении научных статей, но и при выполнении прикладных работ. Поэтому учет цитирований в ограниченном списке научных журналов всегда уменьшает реальное использование конкретной научной публикации. При этом полностью игнорируется масса публикаций, основная по своему воздействию на развитие науки и техники. Тем не менее, следует исключить дословное повторение текстов научных публикаций.

Формальный подход к проведению оценки наукометрических показателей сопряжен с рядом проблем. Первая проблема использования наукометрических показателей в отечественной науке связана с тем, что в то время как наукометрические показатели легко вычислить, велик риск их неадекватного использования в качестве единственного критерия оценки многогранной научно-исследовательской деятельности ученого. Вторая проблема связана с тем, что использование наукометрических показателей в качестве критериев оценки научной деятельности провоцирует ученых к стремлению увеличить значения этих показателей всеми доступными способами.

Основополагающая идея наукометрических рейтингов основана на поверхностном взгляде на процесс получения научного результата. Следует отказаться от практики использования при оценке вклада ученого в науку различных искусственных показателей, ориентирующихся на краткую ретроспективу его деятельности, заменив совокупностью взаимоувязанных интегральных оценок, опирающихся на среднесрочную и долгосрочную ретроспективу научной деятельности.

В связи с этим целесообразно разрабатывать какую-либо балльную оценку различных типов публикаций, которая могла бы учитывать и качественную составляющую опубликованных статей. Возможно составление раздельных рейтингов по разным видам публикаций с последующим усреднением их результатов или учету каждого из рейтингов с разными весовыми коэффициентами, как это делается при наличии разных показателей в рейтингах университетов. Для монографий и глав в монографиях баллы должны рассчитываться с учетом популярности издательства и вклада автора, а также экспертного мнения, выражающегося в наличии рецензий на монографии в ведущих научных журналах или поддержку издания авторитетными научными фондами.

Если мы понимаем под эффективностью научной деятельности отношение результата к затратам на его достижение, то в случае фундаментальных исследований логичным будет расчет количества публикаций на финансирование исследования, результатом которого стали данные публикации.

Необходимо регулярно и подробно обсуждать научные результаты, полученные за отчетный срок. Оценка деятельности научных работников и коллективов должна даваться после тщательной экспертизы и публичного обсуждения полученных научных результатов. Наукометрические показатели, рассчитанные по числу публикаций и цитирований в научных журналах, могут играть лишь вспомогательную роль.

Используемые и не использованные ранее в научных организациях оценочные системы, различные по давности, жесткости и комплексности, в свою очередь также включили в себя вышеупомянутые индексы, а в некоторых – оценка индивидуальной эффективности сотрудников стала строиться исключительно на данных наукометрических показателей. Отличительной особенностью используемых показателей является их фактическая ретроспективность даже при преобладании в оценочных процедурах данных последних лет [11].

Система оценки, стимулирующая к высоким достижениям, предусматривающая содержательную экспертизу, опирающуюся на мировые индексы, способна быть внедрена решительными действиями [2]. Оценка деятельности научных работников и коллективов должна даваться в результате тщательной экспертизы и публичного обсуждения полученных на-

учных результатов [12]. Эксперты, оценивающие успехи ученых для решения финансовых вопросов должны являться специалистами, чьи собственные достижения признаны в научном сообществе [2].

Для объективной экспертизы недостаточно использовать только формальные показатели, так же недостаточно использовать и только экспертные оценки, причем во всех областях наук [13]. Для того чтобы при экспертизе научных проектов, журналов, при оценке эффективности работы научно-исследовательских институтов, лабораторий и отдельных ученых за заданный период времени, соединить достоинства формальных показателей и экспертных оценок, необходимо каждую заявку оценивать множеством независимых экспертов. Каждый эксперт ставит баллы каждому показателю, а затем баллы экспертов суммируются с весами, задающими относительную важность показателей. Остается только упорядочить заявки по убыванию суммы баллов и выбрать для продления первые N заявок, где N определяется либо заданным числом победителей, либо заданным бюджетом конкурса.

Однако экспертный подход имеет целый ряд недостатков. Во-первых, никакой эксперт не может быть абсолютно объективным, т.е. его оценка имеет погрешность и отличается от истинной в предположении, что таковая существует. Во-вторых, еще более опасными могут быть погрешности, порожденные необъективностью отдельных экспертов.

Заметим, что в существующей теории экспертных оценок отмечается, что точность выводов повышается, если критериев достаточно много, причем коррелированность критериев не играет существенной роли. На практике введение большого числа критериев создает проблемы сбора данных, выбора весов и поэтому не может быть рекомендовано.

Предлагается следующий вариант решения задачи об определении победителя при выдаче денежного гранта на основе экспертной оценки нескольких простых формальных критериев:

- уровень научных публикаций;
- опыт ведущего ученого по руководству научным коллективом;
- опыт и возможности ведущего ученого по подготовке научных и педагогических кадров;
- актуальность планируемых научных исследований;
- достижимость заявленных результатов в предложенные сроки;
- соответствие запрашиваемого финансирования поставленным целям, качество проработки сметы проекта;
- перспективный облик лаборатории, создаваемой в организации в рамках проекта в будущем;
- публикационная активность коллектива участников заявляемого проекта;
- имеющаяся у коллектива участников заявляемого проекта научная инфраструктура;
- адекватность принимаемых организацией обязательств по созданию лаборатории;
- кадровый состав организации;
- роль лаборатории в решении задач организации по ее инновационному развитию.

Ученые и их научные исследования имеют дело с идеями и понятиями. Сложно оценить производительность ученого и оригинальность его идей непосредственно. Однако публикации научных статей и научные журналы могут быть взяты за основу в качестве точной меры для анализа научно-исследовательской деятельности, креативности и воздействия людей и государств на ход научного прогресса [14].

НАБОР ПОКАЗАТЕЛЕЙ ДЛЯ ВКЛЮЧЕНИЯ В ГОСУДАРСТВЕННУЮ НАУКОМЕТРИЧЕСКУЮ СИСТЕМУ

Проблема оценки результатов научной деятельности на базе наукометрического анализа к настоящему времени весьма обострилась. Об этом свидетельствуют широко представленная в литературе нелицеприятная критика используемых наукометрических методик: индексов цитируемости, импакт-факторов и других. Предлагаемые усовершенствования используемого сейчас "индекса цитируемости" проблему не снимают, поскольку не могут устранить тот факт, что "востребованность" научной публикации и ее "цитирование" – далеко не совпадающие понятия. Выявление методики оценки именно "востребованности" продолжает оставаться актуальным [15].

Выявление ученых, результативно занимающихся научной деятельностью, должно происходить по сравнительно высоким значениям таких индивидуальных показателей, как среднее число ссылок на статью и максимальное число ссылок на статью [7]. Количество публикаций и число цитирований не могут служить индикаторами оценки качества научно-исследовательской деятельности. Практика цитирования, взятая с чисто количественной точки зрения, усложняет и замедляет процесс получения реальной картины эффективности научной деятельности [8].

Прежде всего, при оценке вклада ученого в науку следует отказаться от практики использования различных искусственных показателей, ориентирующихся на краткую ретроспективу его деятельности, заменив их совокупностью взаимоувязанных интегральных оценок, опирающихся на среднесрочную и долгосрочную ретроспективу научной деятельности. Для формирования системы оценки вклада ученого в науку нужны инновационные меры. Каждая из них должна интегрировать его деятельность за несколько последних лет, а не за короткий промежуток времени [15].

Оценку наукометрических показателей возможно осуществлять на основании анализа ключевых слов и аннотаций, авторов, тем публикации, списков литературы, языка изложения и т.д. Нужно оценивать ссылки не только на научные статьи, но и на другие результаты интеллектуальной деятельности (патенты, программы, ноу-хау и т.д.). ГНС должна содержать в себе справочную систему по патентам и другим объектам интеллектуальной собственности. Повысить эффективность использования индексов цитирования возможно по-разному оценивая публикации: исследовательские тезисы, заключительные отчеты и промежуточные исследовательские отчеты.

Большинство научных статей можно условно разделить на три группы. В первую группу попадают так называемые исследовательские тезисы – публикации, в которых коротко, без доказательств сообщается о полученных результатах. Во вторую группу попадают исследовательские отчеты – публикации, в которых авторы достаточно регулярно отражают промежуточные итоги своих исследований. В третью группу следует отнести статьи, написанные как итоговые отчеты программных исследований. Под исследовательской программой следует понимать научную цель, которую автор ставит перед собой на ближайшие несколько лет. Каждая статья должна содержать четкое и краткое изложение этой цели и указание, какая часть программы реализована в данной статье.

Решающая роль в оценке публикации должна принадлежать специальной комиссии, которая проводит градацию имеющихся индексов, используя соответствующие "веса", наибольший из которых соответствует публикациям третьей категории, а наименьший – публикациям второй категории. Тем самым будет адекватно отражен вклад в научное творчество исследовательских отчетов, полный отказ от которых невозможен и нецелесообразен. При этом не будет утеряна ведущая роль публикаций – отчетов, завершающих исследовательские программы. Высокая градация тезисов связана с "дороговизной" места в известном журнале [16].

При ранжировании вклада ученых в науку надо следовать иерархической классификации наук. При этом нужно обратить внимание на следующие стороны научно-исследовательского процесса:

- защита диссертаций (эта характеристика вполне адекватна при условии правильного функционирования системы оценки диссертаций);
- научные публикации;
- полученные гранты;
- уровень признания.

Уровень признания измеряется относительно просто оцениваемыми характеристиками:

- статус конференций, на которых научный работник делал пленарные доклады;
- число и уровень визитов, оплаченных принимающей стороной;
- полученные награды и премии;
- упоминания в прессе;
- условия труда.

Чем выше научный вклад исследователя и его группы, тем больше должно защищаться диссертаций, тем серьезнее должны быть публикации, тем выше качество представляемых проектов, и, конечно, тем выше уровень признания научной общественностью [17].

Предлагается использовать унифицированные шаблоны наукометрических показателей в виде спецификаций, которые необходимо заполнить значениями из списков и определить параметры. Например, обобщенный индекс цитирования, обобщенный импакт-фактор и т.д. Если использовать шаблоны, то из них должны получаться индекс Хирша, ***H***- и ***G***-индексы; разные виды импакт-факторов. Одна из причин использования импакт-факторов научных журналов в качестве наукометрического показателя заключается в том, что таким путем оценку научной продуктивности можно проводить с помощью соответствующего программного продукта. Достаточно составить базу данных из списков литературных ссылок в электронных версиях журналов и формально их обработать [18].

Использование импакт-фактора (***IF***) журналов регулярно растет не только среди библиометрического сообщества, но и также между учеными и разработчиками государственной научной политики. Одной из причин, объясняющей успех импакт-фактора, является его большая доступность, поскольку разработчики импакт-фактора обеспечили его расчет для большинства международных и всеми признанных журналов. Для каждого отдельного журнала, импакт-фактор используется в качестве косвенного показателя качества и ожидаемого влияния каждой из статей, опубликованной в нем.

Тем не менее, на значение импакт-фактора влияют различные факторы, такие как тематика, тип документов или длина измерительного окна цитирования. Фундаментальные исследования показывают более высокие значения импакт-фактора, чем исследования в области прикладной науки. Как следствие, импакт-фактор следует применять с осторожностью, сравнения его значений должны быть ограничены сопоставимыми единицами.

Используются два различных показателя: показатель активности и относительный импакт-фактор. Индекс активности (***IA***) – это соотношение между долей той или иной конкретной области в данном издании и доли конкретной области в общем числе публикаций. Индекс активности выше 1 указывает на высокую активность, в то время как индекс ниже 1 отражает активность ниже средней. Аналогично, относительный импакт-фактор (***RIF***) определенной области в данной дисциплине рассчитывается как соотношение между средними ожидаемыми по журналу и в среднем по дисциплине.

Расчет ***IF*** – это очень трудоемкий, дорогостоящий и отнимающий много времени процесс. Широкое использование импакт-фактора в этом случае продуцируют некоторые негативные последствия. Злоупотребление и неправильное использование являются основными причинами таких последствий. Приоритет международных и национальных журналов на повестке дня среди ученых; приоритет международных и национальных субъектов исследования и рассмотрение "медленно" меняющихся дисциплин – вот некоторые из наблюдаемых последствий [19]. Импакт-фактор журнала отражает среднее качество статей, в нем опубликованных. Другого подхода к количественной оценке качества статьи до появления сведений о ее цитировании кроме экспертной оценки пока не существует [9].

В соответствии с предложенными принципами отбора наукометрических показателей, для включения в состав ГНС рекомендованы показатели, представленные в таблице.

Наукометрические показатели, предлагаемые для включения в государственную наукометрическую систему

Название наукометрического показателя	Общепринятое обозначение наукометрического показателя (или сокращенное название)	Оценка наукометрического показателя		Назначение
		Методика	Базовые показатели	
Индекс Хирша	***H***	Исследователь имеет индекс ***H***, если ***H*** из его ***N*** статей цитируются по максимуму ***H*** раз	Количество ***N*** опубликованных статей одним исследователем	Оценка продуктивности научного исследователя
Индекс Хирша организации	H_o	Организация имеет индекс H_o, если H_o из ее ***N*** статей цитируются по максимуму H_o раз	Количество ***N*** статей, опубликованных организацией	Оценка продуктивности научной организации
Индекс Eggh'a	***G***	Для данного множества статей, отсортированного в порядке убывания количества цитирований, которые получили эти статьи, ***G***-индекс – это наибольшее число такое, что ***G*** самых цитируемых статей получили (суммарно) не менее G^2 цитирований	Количество ***G*** научных статей	Оценка продуктивности исследователя на основании библиометрических показателей
Индекс Космульского-Пратхала	***I***	Научная организация имеет индекс ***I***, если не менее ***I*** ученых из этой организации имеют ***H*** – индекс не менее ***I***	Индекс Хирша	Оценка публикационной активности научной организации на основании анализа библиометрических показателей
Число самоцитирований	***NSC***	Оценивается методом простого подсчета	Элементарный показатель	Оценка таких личных качеств автора, как недостаточная добросовестность в связи со склонностью к автоплагиату или положительная основательность в написании собственных трудов

Название наукометрического показателя	Общепринятое обозначение наукометрического показателя (или сокращенное название)	Оценка наукометрического показателя		Назначение
		Методика	Базовые показатели	
Число публикаций отечественных ученых в зарубежных журналах	***NPFJ***	Оценивается методом простого подсчета	Элементарный показатель	Оценка продуктивности деятельности научного исследователя
Число публикаций в журналах из списка ВАК	***NPRJV***	Оценивается методом простого подсчета	Элементарный показатель	Оценка продуктивности научного исследователя и его вклада в отечественную науку
Число публикаций в отечественных переводных журналах	***NPRTJ***	Оценивается методом простого подсчета	Элементарный показатель	Оценка продуктивности деятельности научного исследователя и его вклада в отечественную науку
Число публикаций в журналах с ненулевым импакт-фактором	***NPUIFJ***	Оценивается методом простого подсчета	Элементарный показатель	Оценка продуктивности деятельности и качественного уровня публикаций научного исследователя
Год первой публикации	***YFSP***	Оценивается на основании имеющихся библиографических данных	Элементарный показатель	Оценка приоритета и исторического первенства автора в научных исследованиях
Число публикаций по результатам выполненных отечественными учеными исследований и прикладных разработок в ведущих научных журналах	***NPRRJ***	Оценивается методом простого подсчета	Элементарный показатель	Оценка продуктивности научно-исследовательских работ по государственным программам научных исследований
Число патентных заявок, поданных отечественными учеными по результатам завершения научных исследований и прикладных разработок	***NPD***	***NR + ND***	***NR*** – число поданных патентных заявок по результатам исследований; ***ND*** – число поданных заявок по результатам разработок	Оценка продуктивности прикладных научных исследований по программе

Название наукометрического показателя	Общепринятое обозначение наукометрического показателя (или сокращенное название)	Оценка наукометрического показателя		Назначение
		Методика	Базовые показатели	
Число патентов (в том числе признанных международными) на созданные отечественными учеными результаты интеллектуальной деятельности, включая свидетельства о регистрации программ для ЭВМ и баз данных	***NPRIAF***	***NPO + NPF***	***NPO*** – число отечественных патентов на результаты интеллектуальной деятельности, ***NPF*** – число зарубежных патентов на результаты интеллектуальной деятельности за ***x***-й год	Оценка продуктивности отечественных и зарубежных прикладных научных исследований в рамках больших программ научных исследований
Количество проведенных конференций, симпозиумов и выставок в области науки	***NCSE***	***NCON + NSIN + NE***	***NCON*** – количество конференций; ***NSIN*** – количество симпозиумов, ***NE*** – количество выставок за ***x***-й год	Оценка продуктивности, новизны, актуальности и признания научным сообществом результатов, достигнутых в рамках исследований, выполненных по программе
Суммарное количество публикаций отечественных ученых в средствах массовой информации и периодических научных изданиях	***NPMMSJ***	***NMM + NPP***	***NMM*** – количество публикаций в средствах массовой информации; ***NPP*** – количество публикаций в периодических научных изданиях за ***x***-й год	Оценка продуктивности научных исследований по программе, а также интенсивности их популяризации и информационного освещения
Число публикаций отечественных ученых в отечественных и зарубежных средствах массовой информации	***NPRFMM***	***NPMMR + NPMMF***	***NPMMR*** – количество публикаций в отечественных средствах массовой информации; ***NPMMF*** – количество публикаций в иностранных средствах массовой информации	Оценка интенсивности популяризации, освещения и обсуждения результатов исследований, выполненных в рамках программы проведения научных исследований
Индекс оперативности	***IO***	***CIT J (Y, Y) / PUB J (X)***	***CIT (X, Y)*** – суммарное число цитирований, полученных в году ***Y*** теми статьями журнала ***J***, которые вышли в нем в году ***Y*** ; ***PUBJ (X)*** – суммарное число публикаций в журнале в ***x***-ом году	Оценка продуктивности научного журнала

Название наукометрического показателя	Общепринятое обозначение наукометрического показателя (или сокращенное название)	Оценка наукометрического показателя		Назначение
		Методика	Базовые показатели	
Коэффициент самоцитируемости журнала	***KSCD***	***CIT JJ (X) / ∑ CIR IJ (X)***	***CIT IJ (X)*** – число ссылок на журнал ***I*** из журнала ***J*** в ***x***-ом году	Оценка степени активности журнала
Коэффициент самоцитирования журнала	***KSCJ***	***CIT JJ (X) / ∑ CIR JI (X)***	***CIT IJ (X)*** – число ссылок на журнал ***I*** из журнала ***J*** в ***x***-ом году	Оценка востребованности журнала
Индекс Прайса	***P***	***NLX / NL***	***NLX*** – количество ссылок на литературу, опубликованную за последние ***X*** лет до выхода цитирующей статьи; ***NL*** – общее количество ссылок на статью	Оценка популярности научной статьи
Момент первого цитирования	***MFC***	Оценивается на основании имеющихся библиографических данных	Элементарный показатель	Определяет, когда публикация впервые используется и меняет статус с "невостребованного" на "востребованный"
Коэффициент цитируемости группы авторов в целом по научному журналу	***LJ***	***ANCA / ANCJ***	***ANCA*** – среднее число цитирований на одну статью у группы авторов; ***ANCJ*** – среднее число цитирований статей в журнале, где авторы опубликовались	Оценка соответствия уровня авторов уровню журнала
Коэффициент цитирования группы авторов в целом по научной дисциплине	***LD***	***ANCA / ANCD***	***ANCA*** – среднее число цитирований на одну статью у группы авторов; ***ANCD*** – среднее число цитирований одной статьи по дисциплине, в которой авторы работают	Оценка соответствия уровня авторов уровню научной дисциплины
Индекс активности	***IA***	***PR / PC***	***PR***- доля определенной дисциплины в общем массиве публикаций региона; ***PC*** – доля этой же дисциплины в общем массиве публикаций всей страны	Оценка степени востребованности дисциплины в регионе по отношению к востребованности дисциплины в целом по статье

Название наукометрического показателя	Общепринятое обозначение наукометрического показателя (или сокращенное название)	Оценка наукометрического показателя		Назначение
		Методика	Базовые показатели	
Число ссылок на нежурнальные публикации отечественных ученых, включая отчеты о научно-исследовательских работах и патенты	***NLUJPP***	Оценивается методом простого подсчета	Элементарный показатель	Оценка склонности научных исследований к прикладным областям
Число действующих в нашей стране научно-исследовательских институтов	***NBI***	Оценивается методом простого подсчета	Элементарный показатель	Оценка обеспечения научной деятельности по стране
Число публикаций, осуществленных сотрудниками отечественных научно-исследовательских институтов	***NPBI***	Оценивается методом простого подсчета	Элементарный показатель	Оценка продуктивности основных институтов
Среднее число публикаций, приходящихся на один отечественный научно-исследовательский институт	***ANPOBI***	NPC_i / TNI	***NPCI*** – количество статей, опубликованных ***i***-ым институтом; ***TNI*** – общее количество основных институтов	Оценка продуктивности основных институтов
Число научных публикаций, в которых по крайней мере одним из соавторов является гражданин РФ	***NPCAC***	Оценивается методом простого подсчета	Элементарный показатель	Оценка вклада ученых страны в науку на международном уровне

При оценке показателей следует учитывать различные параметры в виде числовых констант, констант типа "дата", наименований наукометрических баз данных, пороговых значений, векторов с коэффициентами, списков научных журналов, научно-исследовательских организаций и т.д.

Параметры должны выбираться по сложной процедуре, чтобы результаты расчетов всегда были актуальными. Если это необходимо, нужно задать границы для "малых" и "больших" значений наукометрических показателей. Однако трудность состоит в том, что границы значений нельзя определить однозначно, так как каждая предметная область имеет собственную специфику, а все предметные области (и другие похожие моменты) учесть сразу вместе не получится ввиду их чрезвычайной широты. Определяя параметры оценки наукометрических показателей, необходимо уточнить, за какое число прошедших лет оценивается конкретный наукометрический показатель, какие источники данных используются для его оценки, какие типы публикаций учитываются при его формировании и так далее.

Наукометрические показатели должны быть сконструированы так, чтобы стремление ученых, а также журналов и научных организаций повысить свои индивидуальные наукометрические показатели способствовало накоплению инновационного потенциала и инновационному развитию экономики на данном этапе. Список наукометрических показателей должен быть гибким, адаптируемым, расширяемым, дина-

мичным. Состав наукометрических показателей ГНС должен регулярно, один раз в три, пять или семь лет, пересматриваться.

ЗАКЛЮЧЕНИЕ

Несмотря на то, что сегодня наукометрические показатели серьезно критикуются за субъективность, недостоверность извлечения, фальсификацию результатов, которые мешают развитию науки и отнимают у нее ресурсы, в настоящей статье дается обоснование тому, что создание государственной наукометрической системы преследует конкретные цели и решает важные задачи.

Наименование государственного задания, в рамках которого выполнена наша статья, предусматривает разработку принципов построения ГНС. Тему государственного задания следует трактовать как своего рода методологическое обоснование крупного проекта, в котором необходимо получить некие общие принципы, т.е. выполнить начальную часть проектных работ или провести предпроектные исследования.

Техническая реализация проекта по созданию ГНС относится к совершенно другому типу задач, и поэтому должна выполняться в рамках научно-исследовательских работ инженерного характера, но тема нашего исследования в первоначальной формулировке носит теоретический, методологический характер. Именно поэтому проектирование ГНС как информационной системы, продолжающее наше исследование о принципах ее построения, следует вынести в отдельную статью и рассмотреть в дальнейшем.

СПИСОК ЛИТЕРАТУРЫ

1. Wu Y., Pan Y., Zhan Y., Ma Z., Pang J., Guo H., Xu B., Yang Z. China Scientific and Technical Papers and Citations (CSTPC): History, impact and outlook // Scientometrics. – 2004. – № 3. – P. 385–397.
2. Чеботарев П.Ю. Оценка ученых: пейзаж перед битвой // Управление большими системами: сборник трудов. – 2013. – № 44. – С. 506-537.
3. Jin B., Zhang J., Chen D., Zhu X. Development of the "Chinese Scientometric Indicators" (CSI) // Scientometrics. – 2002. – № 1. – P. 145–154.
4. Поляк Б.Т. Наукометрия: кого мы лечим? // Управление большими системами: сборник трудов. – 2013. – № 44. – С. 161-170.
5. Стратегия инновационного развития Российской Федерации на период до 2020 года. – URL: http://economy.gov.ru/minec/activity/sections/ innovations/doc20120210_04 (дата обращения: 15.04.2016).
6. Проект государственной программы Российской Федерации «Развитие науки и технологий» на 2013-2020 годы. – URL: http://минобрнауки.рф/documents/2475/ (дата обращения: 15.04.2016).
7. Михайлов О.В. Размышления об оценке научной деятельности // Управление большими системами: сборник трудов. – 2013. – № 44. – С. 144-160.
8. Мотрошилова Н.В. Реальные факторы научно-исследовательского труда и измерения цитирования // Управление большими системами: сборник трудов. – 2013. – № 44. – С. 453-475.
9. Москалева О.В. Можно ли оценивать труд ученых по библиометрическим показателям? // Управление большими системами: сборник трудов. – 2013. – № 44. – С. 308-331.
10. Чеботарев П.Ю. Наукометрия: как с её помощью лечить, а не калечить? // Управление большими системами: сборник трудов. – 2013. – № 44. – С. 14-31.
11. Воронин А.А. Какая эффективность нужна российской науке // Управление большими системами: сборник трудов. – 2013. – № 44. – С. 56-66.
12. Орлов А.И. Два типа методологических ошибок при управлении научной деятельностью // Управление большими системами: сборник трудов. – 2013. – № 44. – С. 32-54.
13. Фрадков А.Л. Блеск и нищета формальных критериев научной экспертизы // Управление большими системами: сборник трудов. – 2013. – № 44. – С. 346-360.
14. Garfield E. The relationship between international science indicators, Nobel class science, and science mapping in the formation of science policy // Statement before the House of Representatives committee on Science & Technology, Science Policy Task Force. – Washington, DC. –1986. – P. 1–22.
15. Гринченко С.Н. Имеет ли решение задача перманентной оценки вклада учёного в науку? // Управление большими системами: сборник трудов. – 2013. – № 44. – С. 280-291.
16. Деза М.М., Деза Е.И. Несколько замечаний к вопросу об оценке научных публикаций // Управление большими системами: сборник трудов. – 2013. – № 44. – С. 362-365.
17. Миркин Б.Г. О понятии научного вклада и его измерителях. // Управление большими системами: сборник трудов. – 2013. – № 44. – С. 292-307.
18. Орлов А.И. Наукометрия и управление научной деятельностью // Управление большими системами: сборник трудов. – 2013. – № 44. – С. 538-568.
19. Bordons M., Fernandez M.T., Gomez I. Advantages and limitations in the use of impact factor measures for the assessment of research performance in a peripheral country // Scientometrics. – 2002. – № 2. – P. 195–206.

Материал поступил в редакцию 22.04.16.

Сведения об авторе

КАЛАЧИХИН Павел Андреевич – кандидат экономических наук, старший научный сотрудник ВИНИТИ РАН, Москва
e-mail: pakalachikhin@viniti.ru

АВТОМАТИЗАЦИЯ ОБРАБОТКИ ТЕКСТА

УДК 81’322.2 : 001.4

В.А. Яцко

Оценка эффективности метрики хи-квадрат

Проведена оценка эффективности применения метрики хи-квадрат с целью взвешивания терминов текстовых документов. Методика оценки предусматривает выбор и предварительную обработку текстов, представляющих классы C и ~C; составление эталонного словаря и начисление коэффициентов входящим в него терминам; получение коэффициентов χ^2 для терминов текста класса C; вычисление обобщенного коэффициента эффективности по сумме коэффициентов, полученных терминами эталонного словаря. Осуществлен сопоставительный анализ взвешивания по формуле χ^2, формуле отношения шансов (OR) и на основе вероятностных величин. Установлено, что лучший результат получен при взвешивании на основе OR.

Ключевые слова: *взвешивание терминов, хи-квадрат, отношение шансов, оценка эффективности, классификация текстовых документов*

Взвешивание терминов является фундаментальным алгоритмом, который широко применяется в различных направлениях лингвистической информатики [1]. В результате взвешивания каждому термину текстового документа приписывается числовой коэффициент и составляется ранжированный список, отсортированный по нисходящей. Термины с наиболее высокими коэффициентами, находящиеся в верхней части списка, считаются статистически значимыми (salient) и могут использоваться в качестве словаря, отражающего содержание данного документа/документов или/и его/их идентификации. Единицами текста (терминами), которым приписываются весовые коэффициенты в процессе взвешивания, выступают отдельные символы (графемы), словоформы (токены), основы слов (стеммы, леммы), словосочетания (н-грамы), предложения, группы предложений, а также и текст в целом.

Алгоритмы взвешивания можно разделить на два основных вида: интертекстуальные и интратекстуальные. Интратекстуальные алгоритмы выполняются на основе анализа внутренней структуры данного текста. Наиболее простым интратекстуальным алгоритмом взвешивания является вычисление вероятностных величин по формуле

$$P_{(i,j)} = \frac{f_{(ij)}}{\sum_{k=1}^{n} f_{(kj)}}, \qquad (1)$$

где $f_{(ij)}$ – частотность термина *i-го* в документе *j-ом,* а k – множество терминов в документе, $k = \{t_i ... t_m\}$. К интратекстуальным относится предложенный нами алгоритм симметричного взвешивания предложений [2], который может применяться с целью автоматического реферирования документов.

Интертекстуальные алгоритмы предполагают сопоставление распределения терминов в данном тексте с их распределением в другом тексте/текстах. Наиболее известным алгоритмом этого типа является алгоритм TF*IDF [3], в соответствии с которым наибольшие коэффициенты получают термины, часто встречающиеся во входном тексте и имеющие небольшую частотность в текстах, с которыми он сопоставляется. Интертекстуальные алгоритмы в основном применяются для решения классификационных задач, таких как авторская атрибуция, категоризация текстов, распознавание плагиата, фильтрация спама. Весовые коэффициенты, начисляемые в процессе автоматической классификации, указывают на дискриминирующую силу терминов – их способность уникально идентифицировать данный класс. В [4] алгоритмы взвешивания, применяющиеся с целью автоматической классификации, разделяются на контролируемые (supervised) и неконтролируемые (unsupervised), при этом под первыми понимаются алгоритмы, предусматривающие отнесение входного документа к некоторому заранее предопределённому классу или категории. TF*IDF относится к неконтролируемым алгоритмам взвешивания, а наиболее распространенными контролируемыми алгоритмами являются хи-квадрат, прирост информации (information gain), отношение шансов (odds ratio), логарифмическое подо-

бие (log likelihood). Метрика хи-квадрат широко используется в медицине для выявления статистически значимых зависимостей между показателями, например, зависимости артериальной гипертонии от курения [5].

Цель настоящей статьи – показать особенности применения метрики хи-квадрат с целью взвешивания терминов, а также оценить её эффективность, сопоставив с результатами взвешивания на основе простых вероятностных величин и метрики отношения шансов (*OR*).

По формуле хи-квадрат (**2**) проводится взвешивание, в результате которого каждому термину класса *C* приписывается числовой коэффициент, указывающий на его дискриминирующую силу.

$$\chi^2(w_c) = \sum_{j=1}^{n} \frac{\left(O(w_j) - E(w_j)\right)^2}{E(w_j)} , \qquad (2)$$

где O – наблюдаемая частотность термина w, а E – его ожидаемая частотность в *j*-ой ячейке таблицы сопряженности (табл.1).

Таблица 1

Таблица сопряженности

Термин	Класс		Сумма
	C	*~C*	
w	*a*	*b*	*S3*
~w	*c*	*d*	*S4*
Сумма	*S1*	*S2*	*S5*

В данной таблице: *a* – частотность термина *w* в классе *C*; *b* – частотность термина *w* в классе *~C*; *c* – частотность остальных терминов в классе *C* (*c=S1 - a*); *d* – частотность остальных терминов в классе *~C* (*d=S2 - b*). Ожидаемая частотность вычисляется как произведение сумм ряда и колонки, в которых располагается данная ячейка, делённое на общее количество терминов в двух классах по формуле:

$$E(w_j) = \frac{(N_j + N_r) * (N_j + N_{cn})}{S_5} . \qquad (3)$$

Также применяется сокращенная формула, по которой хи-квадрат вычисляется только для ячейки *a*:

$$\chi^2(w_c) = \frac{\left(O(w_a) - E(w_a)\right)^2}{E(w_a)} . \qquad (4)$$

Соответственно,

$$E(w_a) = \frac{S_3 * S_1}{S_5} . \qquad (5)$$

Предлагаемая методика оценки метрики хи-квадрат включает следующие этапы.

1. Выбор текстов, представляющих класс C, и класс *~C*. В качестве входного текста, представляющего класс C (t(C)), из офф-лайновой версии Американского национального корпуса (директория nytimes)[1] был произвольно выбран газетный текст *Weighing the risks of liposuction*. Для класса *~C* из газеты *The New York Times* был выбран текст (t(*~C*)) *Body and mind; the high cost of thinness*[2]. В t(C) описывается конкретный случай операции липосакции, а в t(*~C*) обсуждаются общие проблемы пластической хирургии; родо-видовое соотношение между содержанием двух текстов обеспечивает, с одной стороны, совпадение основной терминологии, а с другой – различие в её распределении по частотностям, что обеспечивает адекватное начисление весовых коэффициентов. Оба текста примерно соответствуют друг другу по размеру, в t(C) количество уникальных слов – 604, токенов – 1418, а в t(*~C*) – 604 и 1450 соответственно (см. табл. 2). Из текстов были удалены стоп-слова, указанные в списке Фокса [6]. Также было выполнено распознавание стемм с помощью стеммера Paice/Husk[3], а результаты стемминга отредактированы вручную. Статистические данные были получены с помощью конкорданса AntConc 3.4.4[4]. В вычислениях учитывалось только распределение стемм.

Таблица 2

Статистические данные двух текстов
(без стоп слов)

Текст	Кол-во уникальных слов	Кол-во токенов	Кол-во стемм
t(C)	438	665	391
t(*~C*)	414	609	369
Всего	852	1274	760

2. Составление эталонного словаря. Для составления эталонного словаря трем экспертам-лингвистам было предложено выбрать из входного текста 20 терминов – основ ключевых слов, отражающих его содержание, и приписать каждому термину коэффициент от 1 до 5 в зависимости от его значимости для данного текста. Далее была установлена степень совпадения терминов в списках, составленных экспертами. Во все три списка вошли семь терминов; восемь терминов вошли в два списка из трех; термины, которые встречались только в одном из списков, не учитывались. Для пятнадцати стемм, совпавших в двух и трёх списках, были вычислены среднеарифметические величины совпавших терминов, при этом к стеммам, встречающимся в двух списках, был применён поправочный коэффициент -1. Полученные средние коэффициенты были нормализованы по формуле

$$K_{norm} = \frac{q(w_i)}{q(w_1)} , \qquad (6)$$

[1] http://www.anc.org/data/anc-second-release/anc-second-release-contents/.

[2] http://www.nytimes.com/1988/02/28/magazine/body-and-mind-the-high-cost-of-thinness.html. Тексты из данной газеты считаются эталонными и включаются в корпуса, в том числе в АНК.

[3] http://www.scientificpsychic.com/paice/paice.html.

[4] http://www.laurenceanthony.net/software.html.

где W_1 – термин с самым большим коэффициентом q. Термину с наивысшим коэффициентом приписывался коэффициент 1. В эталонном словаре наибольший коэффициент (**5**) получили три термина, соответственно, коэффициенты остальных терминов делились на это число, а коэффициент 5 в нормализованном списке преобразовывался в коэффициент 1. Дробные числа округлялись до семи десятичных знаков. В табл. 3 приводятся термины эталонного словаря и их коэффициенты. Заметим, что слово *case,* указанное в эталонном словаре, было удалено из текстов как стоп-слово и в дальнейшем не учитывалось.

3. Для каждой стеммы в t(C) был подсчитан весовой коэффициент по формулам (**2**) и (**4**). Коэффициенты хи-квадрат были нормализованы по формуле (**6**), так же как и коэффициенты терминов из эталонного словаря. Обобщённая эффективность метрики оценивалась по формуле

$$Q(m) = \sum_{i=1}^{14} q(w_{i_r}) + q(w_{i_t}) \ , \qquad (7)$$

где $q(w_{i_r})$ – нормализованный коэффициент термина из эталонного словаря, а $q(w_{i_t})$ – нормализованный коэффициент этого термина в t(C), полученный на основе данной метрики.

В табл. 4 представлены результаты взвешивания терминов по метрике хи-квадрат (по полной и сокращенной формулам) и обобщенная оценка её эффективности. Указывается ранг в списке, ранжированном по сумме коэффициентов χ^2 и эталонного словаря (последняя колонка таблицы).

4. Сопоставление со взвешиванием на основе простых вероятностных величин. Для каждого термина в t(C) был подсчитан вероятностный коэффициент по формуле (**1**). Коэффициенты были нормализованы по формуле (**6**) и подсчитана обобщённая эффективность по формуле (**7**). В табл. 5 представлены результаты взвешивания терминов на основе простых вероятностных величин. Указывается ранг в списке, ранжированном по сумме вероятностных величин и коэффициентов эталонного словаря.

5. Сопоставление со взвешиванием на основе метрики отношения шансов (*OR*) [7]. Вычисление значений *OR* также проводится на основе таблицы сопряженности (см. табл. 1) по формуле

$$OR = \frac{d * a}{c * b} \ . \qquad (8)$$

Вследствие большого разброса числовых значений терминов в текстовых документах мы предлагаем использовать модифицированный вариант формулы:

$$OR = \frac{\log(d * (a+1))}{\log(c * (b+1))} \ . \qquad (9)$$

К числу *b* была добавлена единица, для того чтобы избежать ошибки деления на ноль; для выравнивания результатов была добавлена единица и к числу *a*. В табл. 6 представлены результаты взвешивания терминов на основе метрики *OR*. Указывается ранг в списке, ранжированном по сумме коэффициентов *OR* и эталонного словаря.

Таблица 3

Коэффициенты терминов эталонного словаря

№	Термин		Средний коэффициент	Нормализованный коэффициент
	словоформа	стемма		
Термины, совпавшие в трёх списках				
1	blood	blood	4,3333333	0,8666667
2	clot, clots	clot	4,3333333	0,8666667
3	death, deaths	death	4,6666667	0,9333333
4	Hall	Hall	5	1
5	liposuction, liposuctions	liposuc	5	1
6	risk, risks	risk	3,6666667	0,7333333
7	surgery, surgeries surgeon, surgeons, surgical	surg	5	1
Термины, совпавшие в двух списках				
8	case, cases	cas	1,5	0,3
9	complication, complications	complic	3	0,6
10	compression, compressed, compressing, compresses	compress	2	0,4
11	device, devices	devic	1,5	0,3
12	fatal, fatality, fatalities	fatal	2	0,4
13	health	health	1,5	0,3
14	plastic	plastic	3,5	0,7
15	Tiffany	Tiffany	4	0,8

Таблица 4

Эффективность метрики χ^2

Ранг	Термин	Коэффициент χ^2		K_{norm}		$q(w_{i_r}) + q(w_{i_t})$	
		полная формула	сокращенная формула	полная формула	сокращенная формула	полная формула	сокращенная формула
1	Hall	15,7789725	7,4420474	1	1	2	2
4	liposuc	6,0672260	2,8319722	0,3845134	0,3805367	1,3845134	1,3805367
79	surg	1,4382806	0,6627053	0,0911517	0,0890488	1,0911517	1,0890488
12	clot	3,2179542	1,5225564	0,2039394	0,2045884	1,0706061	1,0712550
28	Tiffany	2,7538532	1,3133025	0,1745268	0,1764706	0,9745268	0,9764706
78	blood	1,5271378	0,7225564	0,0967831	0,0970911	0,9634498	0,9637577
337	death	0,0078439	0,0037054	0,0004971	0,0004979	0,9338304	0,9338312
76	risk	1,7791806	0,8398062	0,1127564	0,1128461	0,8460898	0,8461795
5	compress	5,5207372	2,6266050	0,3498794	0,3529412	0,7498794	0,7529412
315	plastic	0,6892368	0,3253326	0,0436807	0,0437155	0,7436807	0,7437155
11	fatal	3,6746954	1,7510700	0,2328856	0,2352941	0,6328856	0,6352941
316	complic	0,5057900	0,2406401	0,0320547	0,0323352	0,6320547	0,6323352
17	devic	2,7538532	1,3133025	0,1745268	0,1764706	0,4745268	0,4764706
317	health	0,5057900	0,2406401	0,0320547	0,0323352	0,3320547	0,3323352
$Q(\chi^2)$						**12,8292498**	**12,8341713**

Таблица 5

Эффективность взвешивания терминов на основе вероятностных величин

Ранг	Термин	коэффициент *Pi*	K_{norm}	$q(w_{i_r}) + q(w_{i_t})$
1	surg	0,0421053	1	2
2	Hall	0,0255639	0,6071429	1,6071429
9	blood	0,0135338	0,3214286	1,3571429
10	liposuc	0,0135338	0,3214286	1,3214286
4	risk	0,0165414	0,3928571	1,1928571
5	clot	0,0150376	0,3571429	1,1571429
8	plastic	0,0150376	0,3571429	1,0904762
11	death	0,0120301	0,2857143	0,9119048
46	Tiffany	0,0045113	0,1071429	0,9071429
19	complic	0,0060150	0,1428571	0,7428571
13	compress	0,0090226	0,2142857	0,6142857
20	fatal	0,0060150	0,1428571	0,5428571
21	health	0,0060150	0,1428571	0,4428571
30	devic	0,0045113	0,1071429	0,4071429
Q(Pi)				14,2952381

Таблица 6

Эффективность взвешивания терминов на основе *OR*

Ранг	Термин	коэффициент *OR*	K_{norm}	$q(w_{i_r}) + q(w_{i_t})$
1	surg	3,623017	1	2
2	Hall	3,5909401	0,9911462	1,9911462
13	liposuc	3,1503543	0,8695388	1,8695388
12	death	3,1632196	0,8730898	1,8064231
6	clot	3,2900550	0,9080980	1,7747647
8	blood	3,2357221	0,8931014	1,7597681
5	risk	3,3046729	0,9121327	1,6454661
37	Tiffany	2,9362961	0,8104559	1,6104559
7	plastic	3,2580700	0,8992697	1,5992697
20	complic	2,9639972	0,8181018	1,4181018
10	compress	3,1796379	0,8776215	1,2776215
18	fatal	3,0333072	0,8372323	1,2372323
21	health	2,9639972	0,8181018	1,1181018
26	devic	2,9362961	0,8104559	1,1104559
Q(OR)				**22,2183461**

В настоящей статье была описана методика оценки эффективности взвешивания терминов текстового документа по формуле хи-квадрат.

Методика включает следующие этапы.

1. Подбор и обработка текстов, представляющих классы *C* и ~*C*. Как правило, эта задача решается в зависимости от целей конкретного проекта. В [8] к данным классам относились тексты определённой тематики, взятые из фармацевтического журнала, поскольку целью проекта было установление жанрово-тематического состава статей, публикуемых в данном журнале. Также широко используются специально создаваемые тестовые корпуса, такие как Reuters-21578[5]. В нашем проекте с целью удобства ручной обработки экспертами были взяты небольшие газетные тексты из двух эталонных источников: Американского национального корпуса и газеты *The New York Times*.

Предварительная обработка текстов предусматривает сокращение их размерности за счет удаления стоп-слов, а также создание ранжированного списка терминов. Под терминами обычно понимаются словоформы, однако, как мы полагаем, более адекватным является начисление коэффициентов стеммам. Словоформы *surgery, surgeries, surgeon, surgeons, surgical* имеют общую семантическую основу, поэтому целесообразно соотнести их с одной стеммой, которой приписывается коэффициент, представляющий собой сумму частотностей указанных словоформ. То же самое относится и к некоторым другим словоформам (см. табл. 3). Для текстов на русском языке предварительный стемминг является особенно актуальным в силу морфологической развитости и наличия большого количества производных форм.

2. Создание эталонного словаря. С целью повышения достоверности результатов и исключения субъективности экспертной оценки нами учитывались не все термины, отобранные экспертами, а только те, которые совпали во всех трёх списках или в двух из них. Термины, которые вошли только в один список игнорировались. Кроме того, терминам в эталонном словаре были начислены коэффициенты для того, чтобы на заключительном этапе получить обобщённый числовой коэффициент эффективности рассматриваемой метрики. В связи с разнородностью получаемых данных была проведена их нормализация на основе соотношения с первым по рангу термином.

3. Получение коэффициентов по формуле хи-квадрат и оценка её эффективности. С этой целью для каждого термина эталонного словаря находилась сумма его коэффициента, вычисленного по формуле хи-квадрат, и коэффициента, полученного на основе экспертной оценки. Далее находилась сумма коэффициентов всех терминов. Полученное число и представляло эффективность данной метрики.

4. Сопоставление эффективности метрики хи-квадрат и других методов взвешивания терминов. Нами было проведено сопоставление с методом взвешивания на основе вероятностных величин, а также с метрикой OR. Первый из них относится к интратекстуальным методам и не предполагает сопоставления с другими текстами; метрика OR относится к той же группе вероятностных методов, что и хи-квадрат, и вычисления также выполняются на основе таблицы сопряженности. Общая эффективность метрики оценивалась по формуле (**7**) как сумма коэффициентов нормализованных терминов эталонного сло-

[5] http://www.daviddlewis.com/resources/testcollections/reuters21578/

варя и коэффициентов, полученных на основе данной метрики. В результате наибольший коэффициент получила метрика *OR*, которая оказалась примерно на 55% эффективнее взвешивания на основе вероятностных величин и на 73% процента эффективнее, чем взвешивание по формулам хи-квадрат.

Заметим, что вычисления проводились по оригинальной модифицированной формуле *OR*, которая и дала непредвиденно хороший результат. Результаты взвешивания по полной и сокращенной формулам хи-квадрат оказались практически идентичными, что объясняется идентичным распределением по рангам. В этой связи представляется целесообразным использовать сокращенную формулу, которая существенно упрощает вычисления. Формула хи-квадрат даёт намного больший разброс терминов по рангам: от 1 до 337, в то время как вариация рангов при взвешивании на основе вероятностный величин – от 1 до 46, а при применении *OR* – от 1 до 37.

Плохой результат метрики хи-квадрат оказался достаточно ожидаемым, так как она применяется для вычисления порогового уровня с целью сокращения размерности текста [9]. Пороговый уровень (квантиль, Qv) находится на основе количества степеней свободы *df* и уровня значимости α. Количество степеней свободы рассчитывается по формуле $df=(R-1)(M-1)$, где R – количество рядов, M – количество колонок, соответственно, для табл. 1 $df=1$. Общепринятой величиной уровня значимости является $\alpha=0.05$[6]; также может использоваться обратная величина $\alpha'=0.95$. В *MS Excel* по формуле ХИ2.ОБР(α';*df*) с округлением до трёх десятичных знаков можно получить Qv=3,841. Поскольку текст t(C) небольшой по размеру, то при применении этого порогового уровня в сокращенный текст войдут только три термина в случае применения сокращённой формулы и десять терминов в случае применения полной формулы. Можно, однако, предположить, что при большем размере текста применение хи-квадрат будет достаточно эффективным. Сказанное не означает, что метрика хи-квадрат в принципе не может использоваться с целью взвешивания терминов. Это возможно при условии её существенной модификации, что требует проведения дополнительных исследований и экспериментов.

СПИСОК ЛИТЕРАТУРЫ

1. Яцко В.А. Компьютерная лингвистика или лингвистическая информатика? // Научно-техническая информация. Сер. 2. – 2014. – № 5. – С. 1–10.
2. Яцко В.А. Методика симметричного взвешивания предложений // Научно-техническая информация. Сер. 2. – 2016. – № 2. – С. 36–41.
3. Yatsko V. TF*IDF revisited // International journal of computational linguistics and natural language processing. – 2013. – Vol. 2, Issue 6. – P. 385–387. URL:https://www.aaai.org/Papers/AAAI/2006/AAAI06-121.pdf.
4. Lan M., Tan C-L., Low H-B. Proposing a new term weighting scheme for text categorization. – 2006. – URL: https://www.aaai.org/Papers/AAAI/2006/AAAI06-121.pdf.
5. Марапов Д. Критерий хи-квадрат Пирсона. – 2013 – URL: http://medstatistic.ru/ theory/hi_kvadrat.html.
6. Fox C. A stop list for general text // SIGIR forum. – 1989. – Vol. 24, № 1–2. – P. 19–21.
7. McHugh M.L. The odds ratio: calculation, usage, and interpretation // Biochemia Medica. – 2009. – Vol. 19, № 2. – P. 120–126. – URL: http://dx.doi.org/10.11613/BM.2009.011.
8. Oakes M.P., Gaizauskas R., Fowkes H. A method based on the chi-square test for document classification // SIGIR '01 Proceedings of the 24th annual international ACM SIGIR conference on research and development in information retrieval. – New York, 2001. – URL: http://pers-www.wlv.ac.uk/~in4326/old/2001_Oakes_SIGIR.pdf.
9. Debole F., Sebastiani F. Supervised term weighting for automated text categorization. – URL: http://nmis.isti.cnr.it/sebastiani/Publications/NEMIS04.pdf.

Материал поступил в редакцию 26.04.16.

Сведения об авторе

ЯЦКО Вячеслав Александрович – доктор филологических наук, профессор Хакасский государственный университет им. Н.Ф. Катанова, г. Абакан
e-mail: viatcheslav-yatsko@rambler.ru

[6] http://math.hws.edu/javamath/ryan/ChiSquare.html

База данных (БД) ВИНИТИ РАН

Федеральная база отечественных и зарубежных публикаций по естественным, точным и техническим наукам, генерируется с 1981 г., обновляется ежемесячно, пополнение составляет около 1 млн документов в год. Тематическое наполнение соответствует реферативному журналу ВИНИТИ. Для поиска одновременно по всем или нескольким тематическим фрагментам генерируется единая Политематическая БД.

БД ВИНИТИ РАН в сети INTERNET

Сервер ВИНИТИ - http://www.viniti.ru – обеспечивает on-line доступ к Базе данных ВИНИТИ РАН круглосуточно без выходных.

На основе БД ВИНИТИ РАН предоставляются следующие услуги:

- Диалоговый поиск научно-технической информации **в режиме on-line**;
- **Демо-версия**, позволяющая ознакомиться с основными функциями поисковой системы, составом данных, формами представления документов и получить навыки работы с системой;
- **Поисковые эксперты** ВИНИТИ выполнят тематический поиск по разовым или постоянным запросам, а также окажут **консультационные услуги**.

БД ВИНИТИ РАН на CD-ROM

Любые наборы тематических фрагментов БД ВИНИТИ или их разделов за любой период с 1981 г., а также **проблемно-ориентированные выборки** из БД ВИНИТИ по актуальным направлениям научных исследований могут быть предоставлены на договорной основе **в поисковой системе (ИПС) "Сокол"**, работающей под управлением Microsoft Windows и обеспечивающей следующие возможности:

- **Чтение** документов в режиме последовательного просмотра или выборочно по оглавлению за весь период заказанной ретроспективы
- **Поиск** документов по автору, заглавию, источнику, ключевым словам или словосочетаниям, реферату, рубрикам, году издания, стране, языку и т.д. (всего более 20 признаков)
- **Словарь** системы поможет правильно подобрать термины для поиска и выбрать глубину их усечения.
- **Для уточнения поиска** можно дополнительно использовать год издания документа, язык текста документа, рубрики, шифры тематических разделов БД.
- Выполненные **запросы можно сохранять** для их последующего использования и/или редактирования.

125190, г. Москва, ул. Усиевича, 20, БД ВИНИТИ РАН.
Отдел взаимодействия с потребителями – (499) 155-45-25, (499) 152-58-81
E-mail: csbd@viniti.ru, sales@viniti.ru

WWW: http://www.viniti.ru

УВАЖАЕМЫЕ КОЛЛЕГИ!

ВИНИТИ РАН предлагает Вашему вниманию Реферативный Журнал в электронной форме

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

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

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

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

ЭлРЖ в версии Windows Вы можете получить за текущий год с любого номера, а также за предыдущие годы.

Подробную информацию Вы можете получить:

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

Коммерческое управление

Телефон/Факс: 8 (499) 155-45-25, 8 (499) 152-58-81

E-mail: contact@viniti.ru, sales@viniti.ru

Центр (Отдел) научно-информационного обслуживания (ЦНИО) ВИНИТИ РАН

Информационные услуги, предоставляемые ЦНИО ВИНИТИ РАН:

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

ВИНИТИ РАН располагает следующими информационными ресурсами:

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

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

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

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

Проводится индивидуальное обслуживание пользователей в читальном зале ЦНИО ВИНИТИ.

Обращаться в ЦНИО ВИНИТИ:
- адрес: 125190, Россия, г. Москва, ул. Усиевича, 20;
- телефоны: 8(499) 155 -42 -43, 8(499) 155 -42 -17;
- эл. почта cnio@viniti.ru, fdk@viniti.ru;
- факс 8(499) 930 -60 -00 (для ЦНИО).