УДК 004.912

ИСПОЛЬЗОВАНИЕ МЕТОДОВ ТЕМАТИЧЕСКОГО АНАЛИЗА В НАУКОМЕТРИЧЕСКИХ СИСТЕМАХ

А. С. Козицын[1, [0000-0002-8065-9061]], **С. А. Афонин**[2, [0000-0003-3058-9269]], **Д. А. Шачнев**[3, [0000-0002-5940-9180]]

НИИ механики МГУ им. М.В. Ломоносова

[1]alexanderkz@mail.ru, [2]serg@msu.ru, [3]mitya57@gmail.com

Аннотация

Во многих современных наукометрических системах и системах цитирования представлены различные механизмы тематического поиска и тематической фильтрации информации. В большинстве случаев для тематического анализа статей и журналов используется полнотекстовый подход, который имеет ряд ограничений. Использование алгоритмов, основанных на анализе графов как автономно, так и совместно с полнотекстовыми алгоритмами, позволяет устранить эти ограничения и улучшить полноту и точность тематического поиска. Алгоритм, разработанный авторами и представленный в этой работе, использует для анализа тематической близости журналов граф соавторства. Алгоритм нечувствителен к языку журнала и подбирает похожие журналы на разных языках, что сложно реализуемо для алгоритмов, основанных на анализе полнотекстовой информации. Апробация алгоритма проводилась в наукометрической системе ИАС ИСТИНА. В интерфейсе, разработанном для этих целей, пользователь может выбрать один близкий ему по тематике журнал, и система автоматически сформирует подборку журналов, которые могут представлять интерес для пользователя как с точки зрения изучения имеющихся в них материалов, так и с точки зрения публикации собственных статей. В перспективе разработанный алгоритм можно адаптировать для поиска похожих по тематике конференций, сборников публикаций и научных проектов. Наличие такого инструмента увеличит публикационную активность молодых сотрудников, повысит цитируемость статей и цитируемость между журналами. Результаты работы алгоритма определения тематической близости между

журналами, сборниками, конференциями и научными проектами также могут использоваться для построения правил в моделях разграничения доступа к данным на основе онтологий предметной области.

Ключевые слова: *тематическая классификация, библиографические данные, граф соавторства, информационные системы*.

ВВЕДЕНИЕ

Применение современных методов тематического анализа для аналитической обработки больших объемов информации используется в настоящее время практически во всех сферах человеческой деятельности, в том числе, в наукометрии. Результаты тематического анализа научной информации могут использоваться в целях уточнения наукометрических показателей, принятия управленческих решений, информационного поиска и определения правил доступа к информации.

Расчет наукометрических показателей используется для оценки значимости статей (цитируемость), авторитетности журналов (импакт-фактор, h5-индекс, h5-медиана), влияния на научное сообщество отдельных авторов (индекс Хирша и g-индекс), оценки деятельности организаций в целом (i-индекс) [1]. Однако многими авторами отмечается, что характеристики распределения абсолютных числовых значений наукометрических показателей имеют существенную зависимость от анализируемой тематической области [2]. Например, значения индекса цитируемости статей за последние 2 года имеют различную медиану для физики и математики, поскольку математические статьи дольше цитируются, но медленнее «набирают» количество ссылок. Аналогичное несоответствие показателей наблюдается и в журналах в целом. Например, лучшие российские журналы по данным РИНЦ, представленные на странице статистики elibrary.ru/titles_compare.asp, за 2018-й год по состоянию на 20.04.2020 имеют следующие показатели цитируемости по разным рубрикам: физика 9200; биология 4600; математика 3600; механика 1500; информатика 1100. В этой связи проводить сравнение по абсолютным значениям наукометрических показателей статей, журналов или авторов из разных тематических направлений некорректно. Необходимо в таких случаях использовать нормализованную среднюю цитируемость [3] или другие аналогичные показатели, учитывающие тематическую область проводимых исследований. Построение таких нормализованных показателей требует проведения тематической классификации больших объемов научных статей и журналов.

При осуществлении управленческой деятельности использование результатов тематического анализа позволяет оценивать состояние различных направлений исследований, проводить их сравнение с мировым уровнем, выявлять новые тематические направления для определения политики выделения материальных ресурсов для стимулирования научной деятельности. При этом необходимо оценивать не только текущие значения показателей, но и их динамику во времени, а также мировые показатели. Например, уменьшение показателей по определенной тематике исследований на фоне роста этих же показателей в мире может означать отток научных кадров в организации из этой области исследований или устаревание оборудования.

Еще одним важным направлением применения тематического анализа является создание эффективных механизмов для проведения информационного поиска. Объектами поиска могут являться: публикации, журналы, персоны, организации и другие объекты. На основе проведения тематической классификации и кластеризации могут решаться такие актуальные задачи, как поиск опубликованных материалов по заданной тематике, поиск наиболее авторитетных экспертов в определенной предметной области, определение списка журналов для публикации и оценка их значимости, выделение новых тематических направлений в какой-либо области и поиск научных коллективов.

Определение тематических связей между объектами информационной системы [4] также может использоваться для автоматического построения онтологий и определения правил доступа к данным в моделях логического разграничения доступа ABAC (Attribute-Based Access Control) [5], которые в настоящее время в значительной степени потеснили старые модели разграничения доступа: ролевую модель RBAC; мандатную модель MAC и дискреционную модель DAC.

Многие крупные наукометрические системы и системы цитирования имеют инструментарий, позволяющий проводить тематический анализ данных.

1. ИСПОЛЬЗОВАНИЕ ТЕМАТИЧЕСКОГО АНАЛИЗА В СОВРЕМЕННЫХ СИСТЕМАХ

Возможности тематического анализа различных наукометрических систем различаются по типу обрабатываемой информации, видам классификаторов, источникам информации, набору используемых методов классификации и кластеризации.

Проект Web of Science (WoS) для проведения тематического поиска использует индексы по ключевым словам и тематические классификаторы. Индексация по ключевым словам производится с использованием авторских ключевых слов (Author Keywords), которые авторы указывают вручную при добавлении статьи. Также производится индексация по ключевым словам и терминам (KeyWords Plus), автоматически выделенным из названий статей, цитируемых в работе. Индексация по ключевым словам позволяет производить поиск и дополнительную фильтрацию с использованием терминов, заданных пользователем. Для индексации по тематическим классификаторам используется два основных классификатора: одноуровневый классификатор Web of Science Categories для журналов, содержащий 250 категорий, и двухуровневый классификатор статей Research Area по 150 областям науки. Кроме того, используется дополнительный классификатор Essential Science Indicators из 22 категорий. В проекте реализован сервис Manuscript Matcher, который позволяет строить рекомендации по подбору журнала для осуществления публикации по тексту рукописи, предлагаемой к публикации [6].

Проект Google Scholar использует двухуровневый классификатор с 8 элементами первого уровня и 400 элементами второго уровня. Разбиение по темам может использоваться на странице scholar.google.com/citations для фильтрации журналов при показе их показателей (h5-индекса и h5-медианы), что позволяет строить более объективные рейтинги для каждой из тематических областей, заданных в классификаторе (Рис. 1). Тематическая фильтрация возможна только для англоязычных журналов. Для русскоязычных журналов, как и для журналов, издаваемых на других языках, тематическая классификация отсутствует.

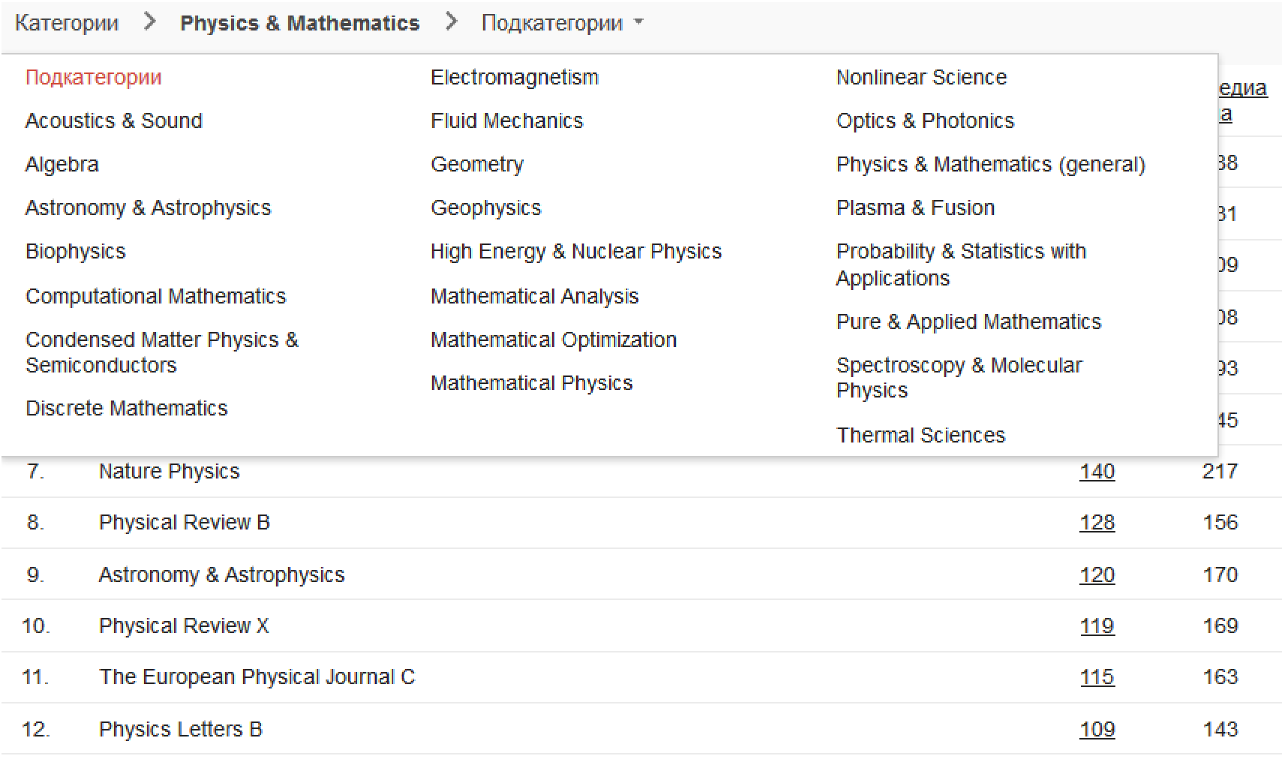

Рис. 1. Интерфейс тематической фильтрации при оценке журнала.

Для корректировки данных Google, в соответствии со своей основной методикой, активно использует взаимодействие с пользователем для сбора информации «снизу–вверх», позволяя авторам создавать собственные страницы со списком статей, фотографией, описаниями интересов (Google Scholar Citations). Добавление статей в профили может производиться автоматизированно (пользователю предлагаются подобранные варианты) или вручную с указанием полных библиографических данных.

Проект Scopus использует двухуровневый классификатор All Science Journal Classification Codes (ASJC), содержащий 4 записи первого уровня и около 350 записей второго уровня для классификации журналов. На странице www.scival.com можно посмотреть распределение журналов по областям и относительные нормированные характеристики для выбранных разделов классификатора, изменение количества публикаций по тематикам с течением времени и другие показатели. Данные доступны при наличии платной подписки.

Проект РИНЦ использует трехуровневый Государственный Рубрикатор НТИ России (ГРНТИ), содержащий около 8 тыс. рубрик (elibrary.ru/rubrics.asp). Тематическую классификацию можно использовать для поиска журналов и статей, а также для фильтрации результатов отбора журналов при выдаче их наукометрических параметров.

Проект Open Academic Graph (OAG), являющийся расширенной версией Microsoft Academic Graph (MAG), содержит 170 млн. статей со ссылками цитирования. Проект не является наукометрической системой или системой цитирования, однако данные проекта могут использоваться для апробации алгоритмов наукометрических систем. Данные можно свободно скачать с сайта проекта www.aminer.org/open-academic-graph.

Кроме перечисленных выше классификаторов коммерческих систем существует целый ряд общепринятых классификаторов, не связанных с какой-то конкретной системой цитирования. На мировом уровне наиболее известным считается трехуровневый классификатор OECD Fields of Science, содержащий более двухсот рубрик, который планировалось использовать, в том числе, в проекте «Карта российской науки». Во многих российских журналах для тематической классификации статей самими авторами используется более подробная Универсальная десятичная классификация (УДК), содержащая более 150 тысяч рубрик. Также для тематической классификации различных научных материалов используются Рубрикатор ВИНИТИ, содержащий более 53 тысяч рубрик, и целый ряд других тематических классификаторов: Классификатор Российского научного фонда (РНФ) [7]; Классификатор Российского фонда фундаментальных исследований(РФФИ) [8]; Международная патентная классификация (МПК) [9], Общероссийский классификатор стандартов (ОКС) [10], Mathematics Subject Classification (MSC) [11]; Journal of Economic Literature Classification (JEL) [12] и другие (scs.viniti.ru/MapService/treeList.aspx). При наличии такого многообразия классификаторов закономерным является появление различных проектов по их согласованию, например, проект по сопоставлению классификаторов Scopus и OECD [13], а также проект ВИНИТИ [14].

Перечисленные выше проекты ставили своей целью разработку систем подсчета показателей цитирования научных публикаций и проведение их тематической классификации по областям науки. Следующим шагом развития стало появление на их основе систем оценки научной деятельности организаций в целом.

Испанский проект SCImago Journal & Country Rank Гранадского университета (или «Атлас науки») оценивает на основе данных Scopus агрегированные данные о научной деятельности в Испании, Португалии и странах Южной Америки. На сайте проекта www.scimagojr.com приводятся показатели не только по научным журналам, но и по странам в целом. Индекс SJR, разработанный авторами проекта, является альтернативой импакт-фактору.

Проект Faculty Scholarly Productivity Index (FSPI) оценивает на основе данных Scopus метрические показатели университетов США. Помимо количества публикаций и показателей цитирования в этом проекте для расчета ранга университета используются данные о полученных наградах и премиях, а также об объемах федерального финансирования исследований. На основе агрегированных данных производится ранжирование более чем 350 университетов.

Проект Times Higher Education (THE) ставит своей задачей оценку университетов всего мира [15]. Разработанный в рамках проекта индекс World University Rankings строится на основе данных о цитируемости WoS, которые составляют 32.5% рейтинга [16]. Помимо этого, учитываются субъективные оценки экспертов, объем финансирования проведенных исследований, привлечение иностранных студентов и преподавателей, а также внедрение разработок вуза в промышленность.

Проект QS World University Rankings [17] оценивает по показателям исследовательской и преподавательской деятельности, соотношению студентов и преподавателей, среднему индексу цитирования в расчете на одного преподавателя, репутации у работодателей, а также количеству иностранных студентов и преподавателей.

Проект Academic Ranking of World Universities (ARWU), который часто называют «Шанхайским рейтингом» (www.shanghairanking.com), учитывает получение выпускниками университета Нобелевских премий, количество опубликованных статей в журналах Nature и Science и показателей цитируемости.

Следует отметить, что проведение подобных сравнений без учета языка преподавания, так же, как и оценка журналов без учета их тематической области, дает не вполне точные результаты [18]. Например, сравнение университетов всего мира по уровню цитируемости только в англоязычных журналах неоспоримо показывает только тот факт, что процент преподавателей и студентов, свободно владеющих английским языком в университетах США, Англии и Канады, значительно выше, чем в России или других не англоязычных странах. Аналогично, сравнение доли иностранных студентов и преподавателей в университетах с английским и японским языками преподавания показывает не столько уровень образования в учебном заведении, сколько количество иностранцев, свободно владеющих данным языком.

Для наукометрических систем, которые ставят своей задачей получение объективных и сбалансированных оценок качества научной продукции, учет области проведения исследования при анализе наукометрических данных, в том числе языка, тематической области и других подобных характеристик, является необходимым требованием при построении объективных наукометрических показателей.

2. ТЕМАТИЧЕСКИЙ АНАЛИЗ С ИСПОЛЬЗОВАНИЕМ ТЕКСТОВОЙ ИНФОРМАЦИИ И КЛАССИФИКАТОРОВ

В процессе разработки и развития наукометрической системы ИСТИНА особое внимание всегда уделялось развитию методов интеллектуального анализа информации, в том числе, методам тематического анализа. Объем данных, обрабатываемых этой системой в настоящее время, значительно уступает мировым системам цитирования, поскольку охватывает всего 28 организаций, 900 тысяч публикаций, 70 тысяч монографий и 13 тысяч патентов. Однако количество типов используемых данных значительно выше. Кроме публикаций и патентов в системе присутствует полная информация о данных по научным проектам (НИР, НИОКР, гранты), докладах на конференциях, диссертациях и дипломах, об участии сотрудников в деятельности различных советов и редколлегиях, получаемых ими премиях и наградах, читаемых учебных курсах и других данных [19]. Кроме того, информация в системе проходит двойную проверку. Основным принципом работы системы является движение информации «снизу-вверх». На первом этапе пользователь, как наиболее заинтересованное лицо, регистрирует в системе все свои работы, которые отображаются на его персональной странице. На втором этапе ответственные сотрудники подразделений подтверждают достоверность данных. Подобный метод сбора информации с использованием создания персональных страниц в настоящий момент использует в проекте Google Scholar Citations корпорация Google, являющаяся лидером на рынке обработки текстовых данных. Но в силу объективных причин в этой системе невозможно организовать второй этап верификации.

Одним из наиболее простых способов проведения тематического анализа является использование классификаторов с ручным сопоставлением объектов и тематических классов, в том числе использование тематической классификации карточек журналов. Такой подход используется в Scopus, WoS и РИНЦ.

В наукометрической системе ИСТИНА на начальном этапе для анализа активности сотрудников и организаций на различных тематических направлениях также были реализованы методы анализа с использованием рубрикации журналов по различным статическим рубрикаторам. С использованием интерактивного интерфейса на странице статистики организации [20] представлены данные о распределении числа статей, цитируемости WoS, числа авторов и других агрегированных характеристик по рубрикам Scopus и ГРНТИ. Данные могут предоставляться как по отдельным подразделениям, так и по организации в целом с возможностью фильтрации по году публикации, преодолению порогового значения анализируемого показателя и принадлежности к группе журналов: журналы из Scopus, журналы из Top25, журналы из ВАК, сборники статей и другие. При этом можно отдельно указывать метрику фильтрации по пороговому значению и метрику для отображения на диаграмме. Например, можно производить фильтрацию по количеству статей, отображать число ссылок на статью.

Возможно проведение анализа как на уровне организации в целом, так и на уровне каждого подразделения отдельно. Следует отметить, что выбор уровня агрегации особенно полезен с учетом неоднозначности определения подразделения для каждой отдельно взятой публикации. В крупных научных организациях большое количество статей публикуется в соавторстве сотрудниками разных подразделений. При традиционном способе подсчета агрегированные данные по отдельным подразделениям считаются независимо, а потом суммируются. В этом случае совместные статьи учитываются несколько раз, что приводит к искажению итоговых показателей. Использование возможности агрегирования исходных данных как на уровне организации (Рис. 2), так и на уровне подразделения (Рис. 3) делает такие оценки более точными и объективными.

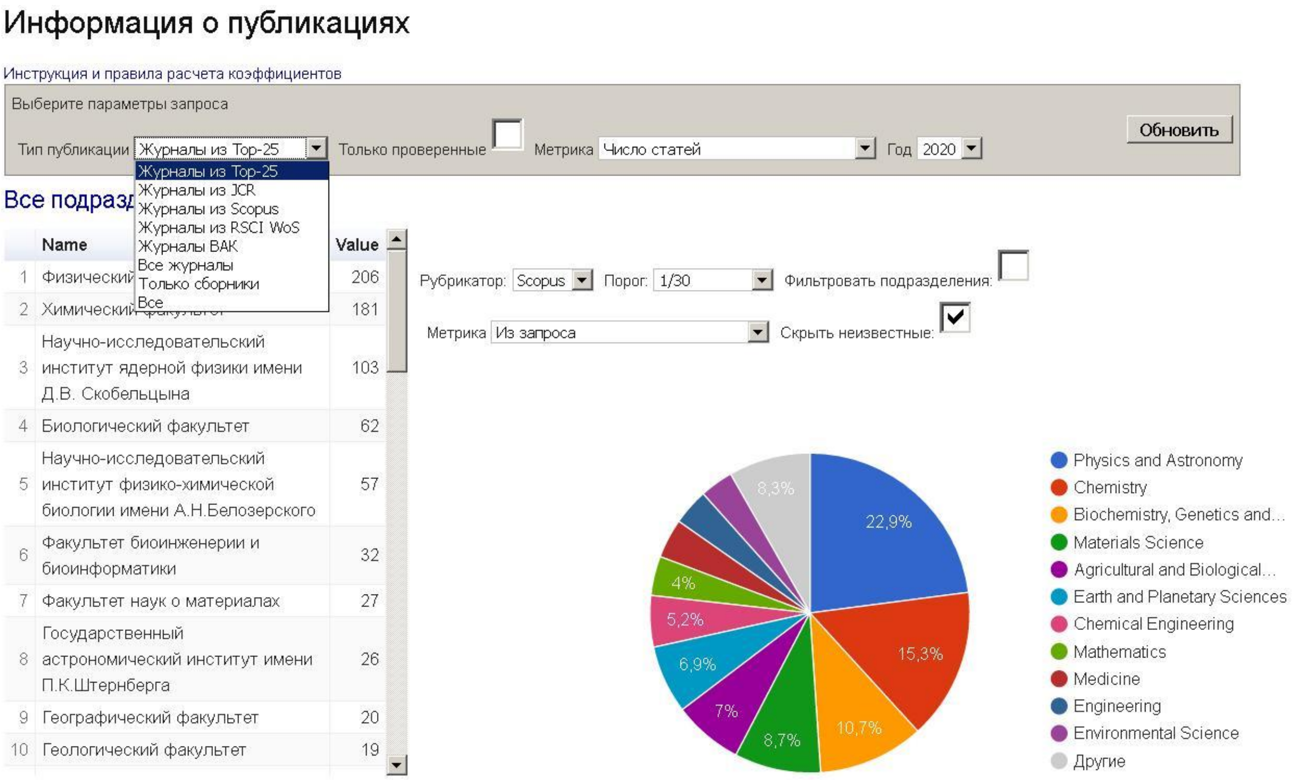

Рис. 2. Интерфейс для анализа распределения числа статей по рубрикам организации.

Подобный подход предоставляет пользователю возможность оценить степень публикационной активности сотрудников в различных тематических областях. Однако он не позволяет анализировать информацию с достаточной степенью детализации. Рубрикатор является статическим, и дополнительная детализация внутри одной рубрики невозможна.

Вторым возможным подходом являются определение тематики и поиск информации по ключевым словам, аннотациям или полным текстам статей. Ключевые слова могут задаваться авторами работы при ее регистрации в системе или вычисляться в процессе индексации из аннотации, полных текстов статей или списка цитируемой литературы, например, Author Keywords и KeyWords Plus в WoS. Такой подход позволяет в большей степени конкретизировать тематику поиска, которая необходима для таких задач, как выделение новых тематических направлений или поиск информации по конкретной информационной потребности пользователя. Следует отметить, что применение подобного тематического анализа не ограничивается только поиском информации. Например, в работе [21] предлагается использовать тематический анализ для оценки качества журнала.

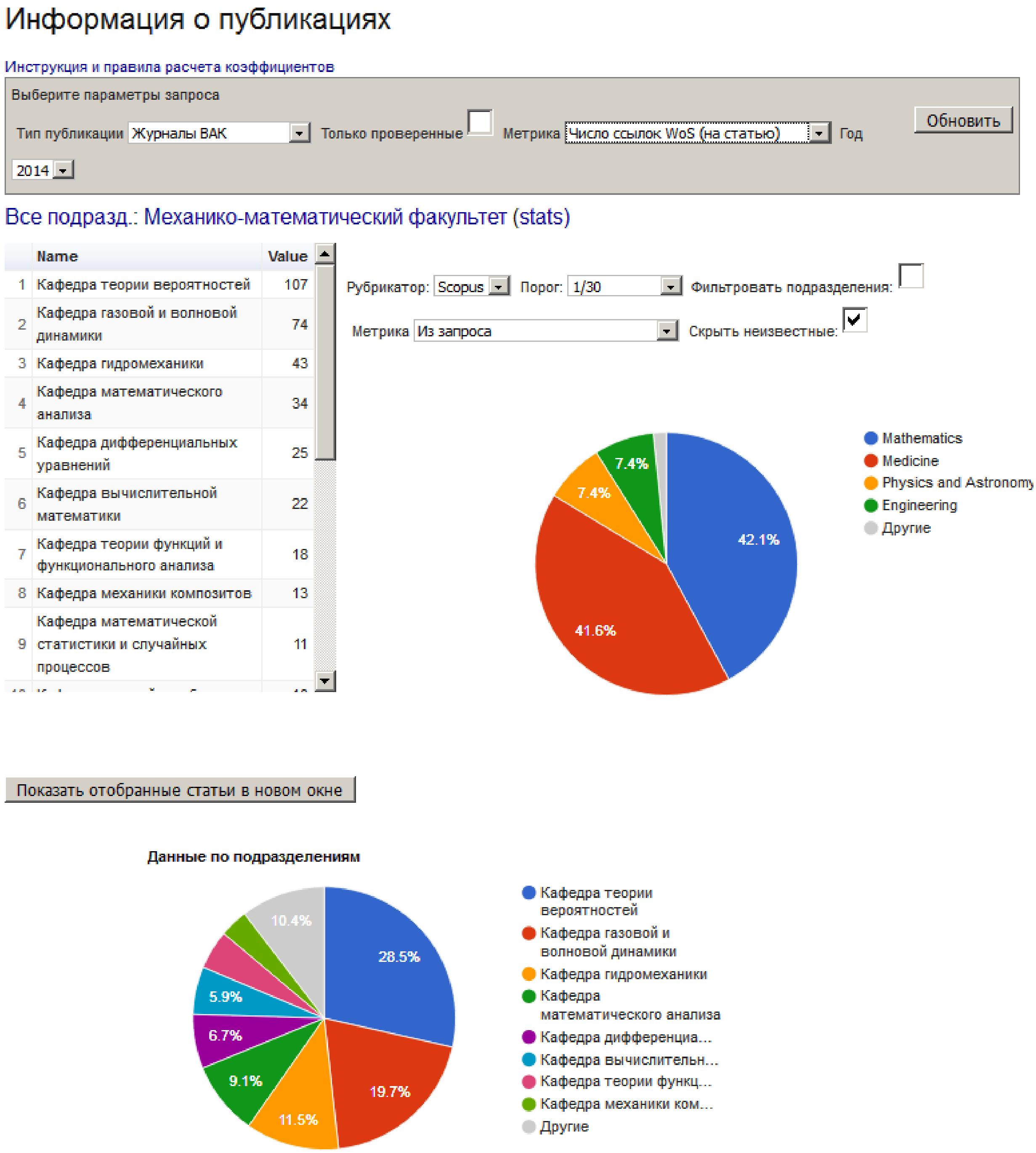

Рис. 3. Интерфейс для анализа распределения данных о цитировании статей подразделения по разным тематическим рубрикам.

Основная гипотеза состоит в том, что в «хороших» научных журналах статьи должны быть посвящены фиксированному набору тематик и эти тематики должны меняться со временем. Таким образом, после обучения алгоритма тематического анализа на обучающем наборе статей из всех анализируемых журналов можно провести тематическую и временную классификацию статей из этих журналов. Качество журнала будет пропорционально точности классификации, с которой содержащиеся в нем статьи были правильно классифицированы по принадлежности к журналу и к временному промежутку публикации.

Основная сложность использования для тематического анализа ключевых слов состоит в ограниченности его набора. При описании статей авторы обычно указывают менее 10 слов. Например, среднее количество ключевых слов, которое авторы указывают при регистрации статей в ИАС ИСТИНА, составляет 3.8. Дополнительным препятствием является субъективность выбора. На первом этапе авторы выделяют из статьи основные понятия, которые, по их оценке, являются значимыми на данный момент. На втором этапе для каждого понятия указывается только одно его отображение на множество ключевых слов без учета возможных синонимов. Таким образом, статьи схожей тематики могут иметь непересекающийся набор ключевых слов, и точность определения их тематического сходства значительно снижается. Вместе с тем, подобный подход к поиску схожих по тематике статей реализован в некоторых системах цитирования. Например, протестировать качество подбора статей при осуществлении поиска по ключевым словам в русскоязычных журналах можно на поисковой странице проекта РИНЦ [22].

Проект WoS предоставляет пользователям сервис Manuscript Matcher подбора журнала для публикации по тексту статьи. Для работы сервис требует предварительной регистрации пользователя. После загрузки заголовка статьи и ее аннотации сервис определяет ключевые слова и ищет соответствие с ключевыми словами журналов. Результат показывается в виде списка журналов с указанием описания журнала, а также списка общих ключевых слов с указанием меры сходства с загруженной статьей. Сервис может быть полезен для авторов, которые используют узкоспециализированные термины, например, в химии, биологии или астрономии. Для более общих тем сопоставление терминов дает не очень точный результат. Например, для статьи "Determining the thematic proximity of scientific journals and conferences using Big Data technologies" лучшими журналами в результатах поиска являются "Scientometrics" и "Journal of the association for information science and technology", однако в top5 попадают журналы "Journal of medical systems" и "Journal of digital imaging" по терминам "create software tools" и "full-text information".

Проект РИНЦ предлагает пользователям сервис поиска похожих статей. Пользователь может выбрать одну из статей, уже проиндексированных в системе, и запросить поиск похожих статей по тематике. Но результаты такого поиска обладают еще меньшей точностью, чем результаты поиска по ключевым словам и работы сервиса Manuscript Matcher. Например, для статьи «Архитектура, методы и средства базовой составляющей системы управления научной информацией «ИСТИНА-НАУКА МГУ»» определяется 14 тысяч близких по тематике статей, и в списке top10 нет ни одной статьи, которая была бы связана с системой, рассматриваемой в статье, или каким-либо аналогом, и только одна статья затрагивает вопросы наукометрии. В результатах поиска по тематическому сходству top3 составляют: "Information technology of software architecture structural synthesis of information system", «Анализ применения asp.net при разработке информационной системы ‘Analysis of the asp.net development information system’», «Общий обзор agris (agricultural research information system)».

Одним из возможных способов улучшения полноты поиска по ключевым словам и разрешения вопросов омонимии явля.тся расширение набора ключевых слов на основе построения связей между ключевыми словами [23], а также использование переводов терминов. В проекте ИСТИНА для автоматизации процесса перевода используются материалы Википедии, а также бесплатные сервисы компании Abbyy. Поиск по ключевым словам используется в качестве первого этапа тематического анализа в разрабатываемых алгоритмах поиска экспертов и подбора журналов, которые апробируются в настоящее время на данных системы ИСТИНА. Результаты исследований, проводимых в этом направлении, нельзя еще использовать для реализации промышленного ПО, однако уже сейчас можно утверждать, что использование только тематического анализа на основе ключевых слов, аннотаций и текстов не позволяет получить удовлетворительный результат классификации. В этой связи в разрабатываемых алгоритмах используется комбинирование методов полнотекстового анализа и методов анализа теории графов, которые анализируют явные или неявные связи между классифицируемыми объектами.

3. ТЕМАТИЧЕСКИЙ АНАЛИЗ С ИСПОЛЬЗОВАНИЕМ СВЯЗЕЙ ОБЪЕКТОВ

Использование связей между объектами (или граф объектов) позволяет дополнить или уточнить данные анализа в случае недостатка информации. Объекты в графе могут быть одного типа, например, статьи и ссылки между статьями, или иметь разный тип, например, сотрудники и их проекты. Целью анализа графа могут быть расширение области поиска или уточнение значимости объектов в существующей области поиска.

Одним из примеров дополнения данных в графе с объектами разного типа является задача поиска экспертов по заданной тематике [24]. Для поиска экспертов в графе авторства определяются объекты, наиболее связанные с экспертами (статьи, монографии, проекты, отчеты и другие), а также степень связи, выделяются ключевые слова объектов, на основе расширенного набора ключевых слов и весов связей графа строится информационный портрет пользователя и оценивается его близость с исходным поисковым запросом. Для апробации алгоритма использовались данные наукометрической системы ИСТИНА. Применение подобных алгоритмов в системах подсчета цитирования затруднено, поскольку граф связей объектов в них содержит только два типа вершин: авторы и публикации. В полноценных наукометрических системах информационный портрет пользователя составляется из большего количества типов объектов, что улучшает качество результатов.

Примером решения задачи уточнения данных на основе связей между аналогичными объектами является алгоритм определения авторства статьей [25], который реализован в системе ИСТИНА. Предполагается, что авторские коллективы обладают определенной устойчивостью, и вероятность публикации двумя авторами нескольких совместных статей гораздо выше, чем написание статьи, в которых одного из авторов заменяет полный однофамилец. В соответствии с этой гипотезой для разрешения неоднозначности при определении авторов статьи среди всех возможных однофамильцев строится граф соавторства и выбирается наиболее связанная компонента.

Используя граф соавторства, можно также решать задачу определения тематической близости журналов без использования данных полнотекстового анализа. Основной гипотезой при реализации этого метода является предположение, что значительная часть авторов публикует статьи в своей предметной области и, следовательно, несколько журналов, в которых публикуется одинаковый набор авторов, близки по тематике. Исходя из этой гипотезы, тематическая близость двух журналов рассчитывается как взвешенная сумма авторов, имеющих публикации в обоих журналах. При этом учитывается не только количество публикаций, сделанных автором, но и позиция автора в библиографическом описании статьи. Вес связи статьи распределяется по всем авторам, но первые авторы имеют больший вес, чем остальные. Формальное описание алгоритма приводится в работе [26]. Основным отличием данного алгоритма от аналогичных алгоритмов, использующих полнотекстовый анализ или анализ по ключевым словам, является нечувствительность к языку журналов и, как следствие, возможность поиска связей журналов на разных языках. Кроме того, алгоритм не нуждается в длительном обучении на больших массивах текстов, показывая при этом достаточно высокую точность 78%.

Дальнейшим развитием описанного в работе [26] алгоритма явились работы по автоматизации расширения области поиска журналов в графе соавторства. Основной предпосылкой является предположение о транзитивности соотношения близости для узкоспециализированных журналов. Если два узкоспециализированных журнала близки по тематике третьему, то они близки между собой. Вместе с тем, обобщение этого правила на все, в том числе общетематические, журналы является неверным. Например, наличие общих авторов у каких либо двух журналов с общетематическим изданием «Известия РАН» не означает взаимной тематической близости исходных двух журналов. В этой связи необходимо использовать математические модели с нормированием весов ребер графа связей журналов [26]. В ходе проведенных исследований было показано, что наилучший результат достигается при проведении нормировки весов ребер с использованием общей суммы исходящих из каждой вершины ребер. После проведения нормировки матрица близости между журналами считается на основе сравнения путей в графе длинной 3. Подобный подход позволяет значительно увеличить полноту тематического поиска. Окончательный результат строится на основе объединения двух списков: наиболее близкие журналы в исходной матрице тематической близости и наиболее близкие журнал в расширенной матрице тематической близости. Объединение этих списков перед показом пользователю позволяет увеличить полноту поиска, не сильно снижая его точность.

Программная реализация алгоритма используется в системе ИСТИНА для предоставления пользователям удобного интерфейса тематического поиска журналов. Для осуществления поиска пользователю необходимо выбрать один известный ему журнал по заданной тематике, найдя его по названию (Рис. 4).

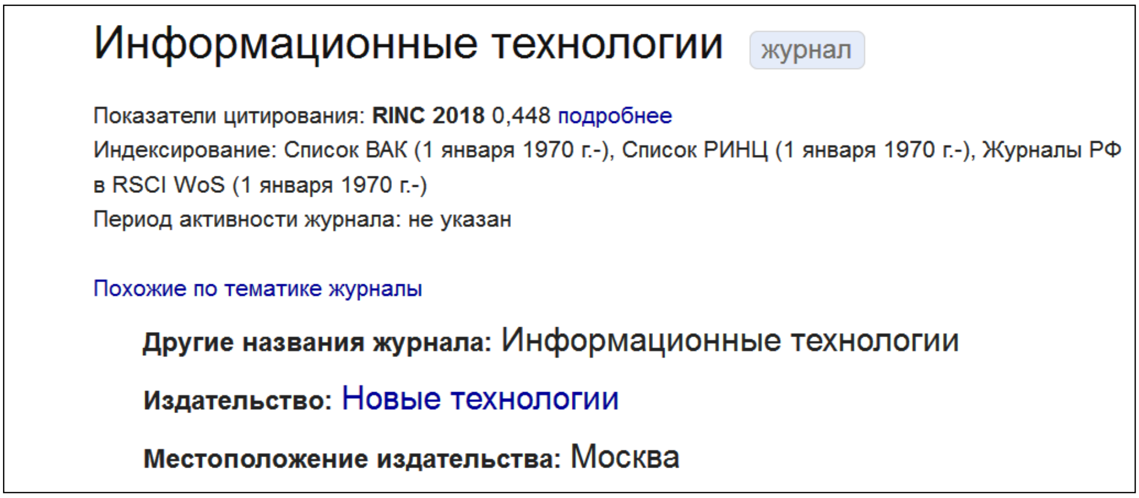

Рис. 4. Карточка журнала.

После этого нужно перейти по ссылке «Похожие по тематике журналы». Для удобства работы в строке каждого журнала в представленном списке указываются оценка его тематического сходства с исходным журналом, различные характеристики цитируемости и количество статей из этого журнала, загруженных в наукометрическую систему (Рис. 5).

Пользователь может перейти на страницу журнала или продолжить перемещение по графу тематических связей журналов, используя ссылки в столбце «Похожие журналы».

Следует отметить, что данный алгоритм может искать тематические связи не только между журналами, но и между другими группами объектов, имеющих авторов. В данном примере алгоритм также ищет конференции, похожие по тематике на заданный журнал.

Show by 10 items Search:

N	Журнал	Вес	Статей за 5 лет	WS	SJR	RINC	Похожие журналы	Похожие конференции
1	Доклады Академии наук	40,7	1061	.195 (1999)	-	1.058 (2018)	журналы	конференции
2	Программная инженерия	39,07	56	-	-	.353 (2017)	журналы	конференции
3	Наука и образование (МГТУ им. Н.Э. Баумана) (электронный журнал)	35,3	17	-	-	-	журналы	конференции
4	Нейрокомпьютеры: разработка, применение	33,38	42	-	-	.341 (2017)	журналы	конференции
5	Программирование	26,62	51	-	-	.685 (2018)	журналы	конференции
6	Programming and Computer Software	21,57	76	.637 (2019)	-	-	журналы	конференции
7	Проблемы информатики	19,27	1	-	-	.138	журналы	конференции

Рис. 5. Интерфейс тематической фильтрации при оценке журнала.

Еще одной важной практической задачей, которая может решаться с использованием описания связей между объектами в наукометрической системе, является задача определения авторитетности экспертов при осуществлении их поиска по тематическому описанию. Для ориентированных графов классическим алгоритмом оценки авторитетности вершин в графе является алгоритм PageRank, который использовался в системе Google для ранжирования результатов поиска. Алгоритм основан на предположении, что входящее ребро в графе подтверждает авторитетность вершины, причем значимость этого подтверждения тем выше, чем выше авторитетность исходящей вершины. В наукометрических системах алгоритм может эффективно использоваться для анализа графа цитируемости. Для анализа неориентированного графа соавторства и других подобных графов в наукометрических системах возможно использование целого ряда других характеристик: степень связности (количество ребер для каждой вершины); степень близости (среднее кратчайшее расстояние до других вершин графа); степень посредничества (количество кратчайших путей между всеми парами вершин, проходящих через заданную вершину); степень влиятельности (степень связности, в которой вклад каждого ребра зависит от степени влиятельности соседней вершины, например, PageRank); кросс-кликовая центральность (число клик, которым принадлежит узел) и другие. Предварительные эксперименты, проведенные на данных системы ИСТИНА, показали, что такой подход может быть достаточно эффективен для использования при ранжировании результата поиска экспертов, автоматического определения устойчивых научных коллективов и других подобных задач.

ЗАКЛЮЧЕНИЕ

Использование алгоритмов тематического анализа для решения целого ряда задач обработки информации в наукометрических системах позволяет создавать удобные сервисы для поиска и обработки информации. Комбинирование полнотекстовых и графовых методов анализа позволяет увеличить точность и полноту представляемых результатов. В настоящий момент в системах научного цитирования такие сервисы представлены недостаточно широко. Научные изыскания на этом направлении, проводимые с использованием данных проекта ИСТИНА, могут позволить получить новые механизмы поиска и обработки наукометрической информации.

Благодарности

Работа выполнена при поддержке Российского фонда фундаментальных исследований, проект 18-07-01055.

СПИСОК ЛИТЕРАТУРЫ

1. *Акоев М.А., Маркусова В.А., Москалева О.В., Писляков В.В.* Руководство по наукометрии: индикаторы развития науки и технологии. Екатеринбург: Издательство Уральского университета, 2014. 248 с.

2. *Орлов А.И.* Наукометрия и управление научной деятельностью // Управление большими системами. Специальный выпуск 44: Наукометрия и экспертиза в управлении наукой. Институт проблем управления им. В.А. Трапезникова РАН. 2013. С. 538–568.

3. *Бричковский В.В.* Наукометрический анализ в информационном обеспечении инновационной деятельности // В мире науки. 2017. № 8(174). С. 64–67.

4. *Афонин С.А., Козицын А.С., Шачнев Д.А.* Программные механизмы агрегации данных, основанные на онтологическом представлении структуры реляционной базы наукометрических данных // Программная инженерия. 2016. Т. 7, №9. С. 408–413.

5. *Afonin S.* Ontology models for access control systems // Proc. of the 3rd International Conference Russian-Pacific Conference on Computer Technology and Applications (RPC). 2018. P. 1–6.

6. Сервис подбора журнала WoS. URL: http://mjl.clarivate.com/home

7. Классификатор РНФ. URL: http://www.rscf.ru/node

8. Классификатор РФФИ. URL: http://www.rfbr.ru/rffi/ru/contest_documents

9. Классификатор МПК. URL: http://www.fips.ru

10. Классификатор ОКС. URL: http://classinform.ru/oks.html

11. Классификатор MSC. URL: http://www.ams.org/msc/

12. Классификатор JEL.
URL: http://www.aeaweb.org/journal/jel_class_system.html

13. Проект по сопоставлению классификаторов Scopus и OECD.
URL: http://report03.metrics.ekt.gr/en/appendixIII

14. Проект по сопоставлению классификаторов ВИНИТИ.
URL: http://scs.viniti.ru/ MapService/mapform.aspx

15. Проект Times Higher Education.
URL: http://www.timeshighereducation.com

16. Индекс World University Rankings.
URL: http://gtmarket.ru/ratings/the-world-university-rankings/info

17. Проект QS World University Rankings. URL: http://www.topuniversities.com

18. *Кинчарова А.В.* Методология мировых рейтингов университетов: анализ и критика // Университетское управление: практика и анализ. 2014. № 2. С. 70–80.

19. Данные проекта ИСТИНА. URL: http://istina.msu.ru/statistics/activity/

20. Статистика организации в проекте ИСТИНА.
URL: http://istina.msu.ru/ statistics/organization/214524/dynamic

21. *Краснов Ф.В.* Сравнительный анализ коллекций научных журналов // Труды СПИИРАН. 2019. Т. 18. С. 767–793.

22. Поиск по ключевым словам в системе РИНЦ.

URL: https://www.elibrary.ru/ querybox.asp

23. *Афонин С.А., Лунев К.В.* Выявление тематических направлений в коллекции наборов ключевых слов // Программная инженерия. 2015. № 2. С. 29–39.

24. *Vasenin V., Lunev K., Afonin S., Shachnev D*. Methods for intelligent data analysis based on keywords and implicit relations: The case of "ISTINA" data analysis system. In Proc. of the International Conference Actual Problems of Systems and Software Engineering (APSSE 2019), IEEE Conference Proceedings, P. 151–155, US, 2019.

25. *Козицын А.С., Афонин С.А.* Разрешение неоднозначностей при определении авторов публикации с использование графов соавторства в больших коллекциях библиографических данных // Программная инженерия. 2017. Т. 8, № 12. С. 556–562.

26. *Козицын А.С., Афонин С.А.* Нахождение скрытых зависимостей между объектами на основе анализа больших массивов библиографических данных // In Proc. of the International Conference Actual Problems of Systems and Software Engineering (APSSE 2019), IEEE Conference Proceedings. 2019. P. 320–328.

THE USE OF THEMATIC ANALYSIS METHODS IN SCIENTOMETRIC SYSTEMS

A. S. Kozitsin, S. A. Afonin, D. A. Shachnev

Institute of Mechanics Lomonosov Moscow State University

alexanderkz@mail.ru, serg@msu.ru, mitya57@gmail.com

Abstract

Modern scientometric systems and citation systems use various mechanisms of thematic search and thematic filtering of information. In most cases, a full-text approach is used for thematic analysis of articles and journals, which has a number of limitations. The use of algorithms based on graph analysis, both independently and in conjunction with full-text algorithms, eliminates these limitations and improves the completeness and accuracy of subject search. The algorithm developed by the authors and presented in this work uses the co-authorship graph to analyze the thematic proximity of journals. The algorithm is insensitive to the language of the journal and selects similar journals in different languages, which is difficult to implement for algorithms based on the analysis of full-text information. The algorithm was tested in the scientometric system IAS ISTINA. In the interface developed for these purposes, the user can select one journal that is close to him on the subject, and the system will automatically generate a selection of journals that may be of interest to the user both in terms of studying the materials available in them and in terms of publishing his own articles. In the future, the developed algorithm can be adapted to search for similar conferences, collections of publications and scientific projects. The presence of such a tool will increase the publication activity of young employees, increase the citation rate of articles and the citation rate between journals. The results of the algorithm for determining thematic proximity between journals, collections, conferences and scientific projects can also be used to build rules in models of differentiating access to data based on domain ontologies.

Keywords: *thematic classification, bibliographic data, co-authorship graph, information systems.*

REFERENCES

1. *Akoev M.A., Markusova V.A., Moskaleva O.V., Pisliakov V.V.* Rukovodstvo po naukometrii: indikatory razvitiia nauki i tekhnologii. Ekaterinburg: Izdatelstvo Uralskogo universiteta, 2014. 248 s.

2. *Orlov A.I.* Naukometriia i upravlenie nauchnoi deiatelnostiu // Upravlenie bolshimi sistemami. Spetsialnyi vypusk 44: Naukometriia i ekspertiza v upravlenii naukoi. Institut problem upravleniia im. V. A. Trapeznikova RAN. 2013. S. 538–568.

3. *Brichkovskii V.V.* Naukometricheskii analiz v informatsionnom obespechenii innovatsionnoi deiatelnosti // V mire nauki. 2017. № 8(174). S. 64–67.

4. *Afonin S.A., Kozitsyn A.S., Shachnev D.A.* Programmnye mekhanizmy agregatsii dannykh, osnovannye na ontologicheskom predstavlenii struktury reliatsionnoi bazy naukometricheskikh dannykh // Programmnaia inzheneriia. 2016. T. 7, №9. S. 408–413.

5. *Afonin S.* Ontology models for access control systems // Proc. of the 3rd International Conference Russian-Pacific Conference on Computer Technology and Applications (RPC). 2018. P. 1–6.

6. Servis podbora zhurnala WoS. URL: http://mjl.clarivate.com/home

7. Klassifikator RNF. URL: http://www.rscf.ru/node

8. Klassifikator RFFI. URL: http://www.rfbr.ru/rffi/ru/contest_documents

9. Klassifikator MPK. URL: http://www.fips.ru

10. Klassifikator OKS. URL: http://classinform.ru/oks.html

11. Klassifikator MSC. http://www.ams.org/msc/

12. Klassifikator JEL.
URL: http://www.aeaweb.org/journal/jel_class_system.html

13. Proekt po sopostavleniiu klassifikatorov Scopus i OECD.
URL: http://report03. metrics.ekt.gr/en/appendixIII

14. Proekt po sopostavleniiu klassifikatorov VINITI.
URL: http://scs.viniti.ru/ MapService/mapform.aspx

15. Proekt Times Higher Education.
URL: http://www.timeshighereducation.com

16. Indeks World University Rankings.
URL: http://gtmarket.ru/ratings/the-world-university-rankings/info

17. Proekt QS World University Rankings. URL: http://www.topuniversities.com

18. *Kincharova A.V.* Metodologiia mirovykh reitingov universitetov: analiz i kritika // Universitetskoe upravlenie: praktika i analiz. 2014. No. 2. S. 70–80.

19. Dannye proekta ISTINA. URL: http://istina.msu.ru/statistics/activity/

20. Statistika organizatsii v proekte ISTINA.
URL: http://istina.msu.ru/statistics/ organization/214524/dynamic

21. *Krasnov F.V.* Sravnitelnyi analiz kollektsii nauchnykh zhurnalov // Trudy SPIIRAN. 2019. T. 18. S. 767–793.

22. Poisk po kliuchevym slovam v sisteme RINTs.
URL: https://www.elibrary.ru/querybox.asp

23. *Afonin S.A., Lunev K.V.* Vyiavlenie tematicheskikh napravlenii v kollektsii naborov kliuchevykh slov // Programmnaia inzheneriia. 2015. № 2. S. 29–39.

24. *Vasenin V., Lunev K., Afonin S., Shachnev D*. Methods for intelligent data analysis based on keywords and implicit relations: The case of "ISTINA" data analysis system // In Proc. of the International Conference Actual Problems of Systems and Software Engineering (APSSE 2019), IEEE Conference Proceedings, 2019. P. 151–155, US, 2019.

25. *Kozitsyn A.S., Afonin S.A.* Razreshenie neodnoznachnostei pri opredelenii avtorov publikatsii s ispolzovanie grafov soavtorstva v bolshikh kollektsiiakh bibliograficheskikh dannykh // Programmnaia inzheneriia. 2017. T. 8, No 12. S. 556–562.

26. *Kozitsyn A.S., Afonin S.A.* Nakhozhdenie skrytykh zavisimostei mezhdu obieektami na osnove analiza bolshikh massivov bibliograficheskikh dannykh // In Proc. of the International Conference Actual Problems of Systems and Software Engineering (APSSE 2019), IEEE Conference Proceedings. 2019. P. 320–328.

СВЕДЕНИЯ ОБ АВТОРАХ

КОЗИЦЫН Александр Сергеевич – ведущий научный сотрудник, к. ф.-м. н., окончил мехмат МГУ им. М.В. Ломоносова. Специалист в области информационного поиска и баз данных.

Alexander Sergeevich KOZITSYN – Leading Researcher, Ph.D., graduated from M.V. Lomonosov Moscow State University. Specialist in the field of information retrieval and database.

email: alexanderkz@mail.ru,
ORCID: 0000-0002-8065-9061

АФОНИН Сергей Александрович – ведущий научный сотрудник, к. ф.-м. н., окончил мехмат МГУ им. М.В. Ломоносова. Специалист в области регулярных языков и информационных систем.

Sergey Alexandrovich AFONIN – Leading Researcher, Ph.D., graduated from M.V. Lomonosov Moscow State University. Specialist in the field of regular languages and information systems.

email: serg@msu.ru， ORCID:0000-0003-3058-9269

Шачнев Дмитрий Алексеевич – программист, окончил мехмат МГУ им. М.В. Ломоносова. Специалист в области информационных систем.

Dmitry Alekseevich SHACHNEV – programmer, graduated from M.V. Lomonosov Moscow State University. Specialist in information systems.

email: mitya57@gmail.com, ORCID: 0000-0002-5940-9180

Материал поступил в редакцию 18 ноября 2020 года