УДК 004.912
DOI: 10.25559/SITITO.16.202001.108-117

Алгоритм разрешения неоднозначности имен авторов в ИАС ИСТИНА

А. С. Козицын*, С. А. Афонин
Московский государственный университет имени М. В. Ломоносова, г. Москва, Россия
119991, Россия, г. Москва, ГСП-1, Ленинские горы, д. 1
* alexanderkz@mail.ru

Аннотация

Актуальной задачей для систем сбора и обработки наукометрических данных является определение авторов публикаций по библиографическим данным. Такая задача является важной как для узкоспециализированных систем анализа цитируемости публикаций с автоматическим сбором информации о публикациях из различных источников, так и для систем с большим охватом научной продукции (публикации, диссертации, патенты, лекции и т.д.) с политикой сбора информации «снизу вверх». Точность определения автора влияет как на результаты наукометрической оценки научной деятельности субъектов, так и на защищенность систем при использовании современных моделей логического разграничения доступа при определении прав пользователей.

В работе описывается разработанный для решения этой задачи алгоритм, который в настоящее время реализован в рамках работ по созданию наукометрической системы ИАС ИСТИНА. Эта система используется с 2012 года для сбора данных о научной деятельности в Московском государственном университете имени М.В. Ломоносова, обрабатывает данные о научных публикациях, педагогической деятельности, НИР, диссертациях, участии в различных советах, получении научных премий и многих других показателях. В настоящий момент ИАС ИСТИНА используется более чем в двадцати организациях.

Основной особенностью разработанного алгоритма является использование для идентификации автора графа соавторства, который строится не только по зарегистрированным в системе более ранним публикациям, но и по другим совместным результатам научной деятельности. В статье описывается структура модуля анализа библиографических данных, который реализован в системе ИАС ИСТИНА, и приводится описание разработанного алгоритма идентификации авторов по библиографическим данным публикации. Результаты тестирования программной реализации алгоритма, приведенные в конце статьи, показывают высокую точность его работы.

Ключевые слова: наукометрия, информационные системы, библиография, автор, граф.

Финансирование: исследование выполнено при финансовой поддержке Российского фонда фундаментальных исследований в рамках научного проекта № 18-07-01055_а «Атрибутная система логического разграничения доступа для информационных систем на основе реляционных баз данных».

Для цитирования: Козицын, А. С. Алгоритм разрешения неоднозначности имен авторов в ИАС ИСТИНА / А. С. Козицын, С. А. Афонин. – DOI 10.25559/SITITO.16.202001.108-117 // Современные информационные технологии и ИТ-образование. – 2020. – Т. 16, № 1. – С. 108-117.

Algorithm for Resolving the Ambiguity of Author Names in IAS ISTINA

A. S. Kozitsin*, S. A. Afonin
Lomonosov Moscow State University, Moscow, Russia
1, Leninskie gory, Moscow 119991, Russia
* alexanderkz@mail.ru

Abstract

An important task for scientometric data collection and processing systems is to identify authors of publications based on bibliographic data. This task is an important citation analysis system for publications. Information in such systems is automatically collected from various sources. It is necessary to automatically process the collected bibliographic data. Such a task is important for scientometric systems that analyze numerous types of scientific products (publications, dissertations, patents, lectures, etc.) too. The accuracy of the author's definition affects the quality of the results of scientometric assessment of the scientific activity of subjects. In addition, the accuracy of determining the author is important for protecting systems that use modern models of logical access control.
The article describes an algorithm developed to solve this problem, which is currently used in the scientometric system IAS ISTINA. This system has been used since 2012 to collect data on scientific activities at Lomonosov Moscow State University. The system processes data on scientific publications, pedagogical activity, research, dissertations, participation in various councils, obtaining scientific awards and many other indicators. Currently, IAS ISTINA is used in more than twenty organizations.
The main feature of this algorithm is the use of a co-authorship graph to identify the author. It is built for publications and other results of scientific activity. In this article id described the structure of the bibliographic data analysis module, which is implemented in the IAS ISTINA system, and the algorithm for identifying authors based on bibliographic data of the publication. The results of testing of the algorithm are presented at the end of the article. It demonstrates the high accuracy of algorithm operation.

Keywords: scientometrics, information systems, bibliography, author, graph.

Funding: the research was carried out with the financial support of the Russian Foundation for Basic Research in the framework of the scientific project No. 18-07-01055_a "Attribute Logical Access Control System for Information Systems Based on Relational Databases".

For citation: Kozitsin A.S., Afonin S.A. Algorithm for Resolving the Ambiguity of Author Names in IAS ISTINA. *Sovremennye informacionnye tehnologii i IT-obrazovanie* = Modern Information Technologies and IT-Education. 2020; 16(1):108-117. DOI: https://doi.org/10.25559/SITITO.16.202001.108-117

Введение

Развитие средств анализа больших данных и тенденция развития открытого представления результатов научной деятельности, в том числе, публикаций, в интернет привело к появлению автоматизированных систем сбора и аналитической обработки наукометрических данных [1]. На первом этапе возникли системы оценки публикационной активности авторов – наукометрические базы данных [2] Web of Science, Scopus, Российский индекс научного цитирования (РИНЦ) на платформе НЭБ eLIBRARY.RU [3], [4], Астрофизическая информационная система Astrophysics Data System, база данных медицинских и биологических публикаций PubMed, библиографические базы данных Chemical Abstracts Plus и Mathematics, международная издательская компания Springer, Международная информационная система по сельскому хозяйству и технологиям International System for Agricultural Science and Technology (Agris), библиографическая база данных по геологии и геонаукам GeoRef и другие. Эти системы отличаются по возможностям анализа и области анализа публикаций. Однако, все системы этого класса анализируют только публикации, и не учитывают другие данные по научной и педагогической деятельности.

Следующим шагом развития стало появление наукометрических систем [5], позволяющих собирать и анализировать большое количество показателей научно-педагогической деятельности, таких как доклады на конференциях, регистрация патенты, участие в научных проектах, преподавание учебных курсов, участие в работе диссертационных советов, научное руководство студентами и аспирантами, получение награды и других. Несмотря на наличие определенных недостатков при проведении объективной оценке научной деятельности ученых [6], использование таких наукометрических систем позволяет получать различные аналитические оценки деятельности организаций и научных коллабораций [7] для принятия управленческих решений [8], [9], [10]. В этой связи, необходимо обеспечивать максимальную точность предоставляемой информации, разрабатывать методы автоматической проверки и коррекции данных при вводе в систему пользователем, использовать для верификации данных внешние источники информации и интеллектуальные алгоритмы агрегации данных [11].

Еще одним важным аспектом, который требует обеспечения качественной проверки вводимой в наукометрическую систему первичной информации о научной и педагогической деятельности, является все более широкое распространение систем обеспечения логического разграничения доступа к данным ABAC (Attribute-Based Access Control) [12], которые приходят на смену традиционным моделям разграничения доступа (ролевая модель RBAC; мандатная модель MAC, дискреционная модель DAC [13], [14], [15]), ориентированным на применение в информационных системах с разделением обязанностей пользователей и фиксированными уровнями доступа. Использование моделей ABAC для определения прав доступа предусматривает использование атрибутов объектов и заданных между ними логических связей, и требует качественного семантического описания предметной области, в том числе автоматических или автоматизированных методов построения онтологий [16], [17]. В рамках модели ABAC разработано множество конкретных решений [18], [19], [20], как узкоспециализированных, в которых вычисление значений атрибутов информационных объектов производится на языке высокого уровня и изменение политики требует изменения исходного кода приложения, так и общих, с возможностью гибкой настройки правил и возможностью автоматической проверки их непротиворечивости [21], [22]. Такое построение оказывается невозможным при наличии большого количества ошибок в исходных данных.

Для уменьшения количества ошибок идентификации авторов при регистрации публикаций в наукометрических системах, и увеличения точности показателей цитирования и наукометрии предпринимаются попытки создания всеобъемлющих баз уникальных идентификаторов авторах научных работ. В 2008 году в компании Thomson Reuters в рамках проекта Web of Science создана система уникальных идентификаторов ResearcherID. Параллельно этой системе идентификации авторов стала развиваться система Scopus Author ID компании Scopus. Идентификация авторов в этих системах основывается на данных библиографических записей, а также на афиллиации авторов. Однако, такой подход имеет некоторые недостатки. Во-первых, авторы указывают в работах различные варианты названия организации. Например, МГУ имени М.В. Ломоносова (МГУ) встречается в работах в следующих написаниях: Lomonosov Moscow State University; Lomonosov State University; Lomonosov Moscow University; Lomonosov MSU; M.V. Lomonosov Moscow State University; M.V. Lomonosov State University; Lomonosov Moscow State University (MSU); Moscow State University named after Lomonosov M.V.; Moscow State Lomonosov University; MV Lomonosov Moscow State University и другие. Во-вторых, в крупных организациях встречаются сотрудники, имеющие одинаковые фамилии и инициалы, и разделить их по принципу принадлежности организации невозможно.

В 2012 году была запущена система ORCID единой идентификации научных исследователей на базе стандарта идентификации авторов ISNI, которая, в отличие от двух предыдущих, уже не являлась проприетарным решением, и в своей основе предполагает самостоятельную регистрацию авторов и явное указание ORCID авторов в публикации. Однако, развитие такого подхода несколько ограничивается необходимостью проведения самостоятельной регистрации пользователями. Для полноценного внедрения таких систем требуется не только мотивировать все научное сообщество для перехода на новый стандарт, но и изменить стандарты библиографических ссылок. Такие процессы в мировом масштабе проходят очень медленно. В 2019 год в системе ORCID зарегистрировано всего 7 млн. авторов научных работ во всем мире. Среди сотрудников МГУ идентификатор ORCID имеют всего 2 тысячи человек. Вместе с тем, в МГУ на настоящий момент по данным системы ИСТИНА зарегистрировано 20 тысяч авторов, опубликовавших 600 тысяч статей. Таким образом, за 7 лет функционирования системы ORCID удалось добиться полноты покрытия около 10%. Полный перехода научного сообщества на новые стандарты публикации может занять несколько десятков лет. В этой связи, необходимо создавать и использовать альтернативные методы обработки библиографических записей, которые бы позволяли с достаточной степенью точности определять идентификатор автора при регистрации публикаций в наукометрических системах.

В настоящей работе приводятся результаты, полученные ав-

торами, при создании наукометрической информационно аналитической системы ИСТИНА (далее ИАС ИСТИНА или система), которая активно используется в настоящий момент в МГУ и еще более чем в 20 научных организациях и ВУЗах, и имеет более 150 тысяч зарегистрированных пользователей [23]. С учетом значительного масштаба системы и относительно большого времени ее эксплуатации, приводимые в настоящей статье данные могут адекватно отражать достоинства и недостатки предлагаемого алгоритма для использования в других научных и педагогических организациях.

В ИАС ИСТИНА предполагается, что движение данных происходит снизу вверх [24]. На нижнем уровне ученый или педагог вводит данные о своей научной деятельности, поскольку именно он лучше всего может точно перечислить свои результаты и наиболее полно может их описать. При этом данные по каждому результату научной деятельности вводятся только один раз одним из соавторов. На более высоких уровнях происходит проверка, подтверждение и агрегация введенных пользователями данных.

Одной из задач, которую необходимо решить при реализации такого подхода, является разработка алгоритмов анализа и обработки библиографических данных с целью создания удобного интерфейса ввода публикаций, полученных из внешних источников или введенных пользователем.

Модуль анализа и обработки библиографических данных

Основной задачей модуля анализа и обработки библиографических данных является получение структурированных данных о публикации по полнотекстовому описанию. Схема работы модуля представлена на рисунке 1. На первом этапе производится очистка данных и разделение их на элементы: название публикации, авторы, название журнала, том, номер, год издания. По полученным данным производится поиск похожих публикаций, которые были загружены в систему ранее. Если такие публикации найдены, пользователю выдается предупреждение. На следующем этапе обработки производится выделение фамилий и инициалов отдельных авторов и поиск «похожих» авторов в базе данных ИАС ИСТИНА для получения их уникального идентификатора (IRID). Следует отметить, что поиск производится с учетом возможных опечаток фамилии или инициалах автора, переводов фамилии автора на другие языки и истории смены фамилии автором.

Р и с. 1. Схема работы модуля анализа и обработки библиографических данных

F i g. 1. Work scheme of the module for analysis and processing of bibliographic data

Важным фактором, влияющим на результаты поиска, является распространенность фамилии автора публикации. По данным ИАС ИСТИНА более 50 тысяч авторов имеют более 10 однофамильцев. В этой связи, использование для идентификации авторов только фамилии и инициалов оказывается недостаточным и необходимо использовать специальные алгоритмы определения наиболее вероятных IRID для каждого автора публикации перед показом результата разбора пользователю. Разработанный авторами настоящей статьи алгоритм использует информацию о графе соавторства публикаций, загруженных в базу данных ранее, и позволяет на его основе ранжировать результаты поиска IRID по вероятности правильного соответствия автору данной публикации.

Алгоритм разрешения неоднозначности определения авторов

Уточнения IRID автора можно осуществлять на основе анализа дополнительной информации, в том числе, устойчивых групп соавторов. Описание таких групп рассчитывается с использованием методов анализа графов на основе графа соавторства, построенного по всем введенным ранее в систему публикациям. Такой подход позволяет не только получать дополнительную информацию об авторах, но и решать более глобальные задачи, такие, как определение авторитетности [25] или проведение тематического анализа без использования знаний о языке документа [26]. В работе [27] представлен обзор существующих в настоящее время алгоритмов классификации документов с использованием описаний авторов с указанием их недостатков для реализации в системе ИСТИНА. К основным недостаткам относятся: требование к наличию информации о email, месте работы автора и списка литературы, которые отсутствуют в библиографическом описании; требование к обучению на разных выборках; составление обучающих выборок с ручной разметкой и другие. Учитывая указанные недостатки, на первом этапе разработки системы ИСТИНА был разработан алгоритм поиска авторов, описанный в работах [27] и [28]. Алгоритм основан на анализе графов соавторства. Для каждого возможного набора авторов производится построение подграфа их соавторства и оценивается вероятность правильности такого набора. Этот алгоритм имеет высокую точность при хорошей обучающей выборке (90%-97%, в зависимости от количества соавторов), однако, имеет большую вычислительную сложность, использует все возможные подмножества соавторов и не использует данные об авторизации пользователей в системе, которые содержат значительное количество информации об авторстве работ. Анализ статистических данных ИАС ИСТИНА показывает, что около 93% публикаций добавляется в систему одним из соавторов. Кроме того, имеется довольно много устойчивых подгрупп авторов, которые часто публикуются в соавторстве, и к ним иногда добавляются другие «случайные» авторы. Использование информации об авторизованном пользователе и об устойчивых подгруппах авторов позволяет значительно улучшить качество определения правильных IRID, уменьшив при этом вычислительную сложность алгоритма за счет поэтапного определения наиболее вероятных авторов вместо оценки различных вариантов подграфа целиком. Такой подход позволил в 10 раз уменьшить время обработки одной статьи. Первые результаты апробации данного подхода были описаны в работе [30].

Входными данными для разработанный авторами статьи алгоритма являются полученные из библиографического описания публикации фамилии и инициалы каждого из авторов, а также граф соавторства. В первой реализации алгоритма [30] для построения графа соавторства использовались только данные из статей, тезисов, книг и проектов. В настоящий момент для охвата всех сфер научной деятельности и улучшения качества распознавания при построении графа соавторства дополнительно анализируются данные по созданию учебных курсов, выступлениях в СМИ, научных достижениях, совместно полученных наградах, патентах и свидетельств о регистрации ПО. Вершинами графа соавторства являются IRID всех зарегистрированных в системе авторов, а вес ребер равен количеству совместных работ, сделанных каждой парой авторов. Для ИАС ИСТИНА такой граф содержит на настоящий момент 340тыс. вершин и 8.2 млн. ребер. В таблице 1 представлено распределение весов ребер графе соавторства.

Таблица 1. **Распределение весов ребер**
Table 1. **Distribution of edge weights**

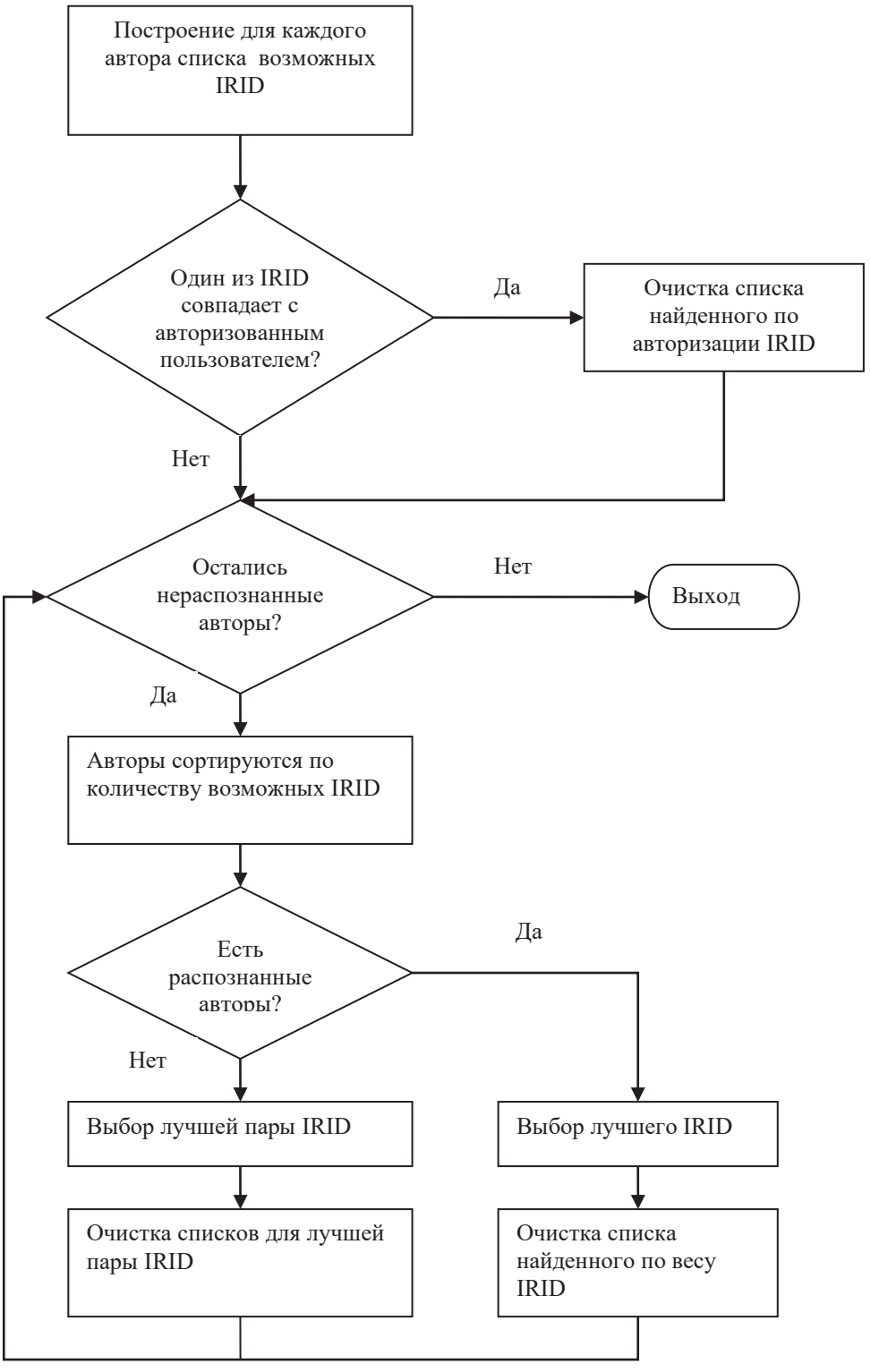
Вес ребра	Количество ребер (тыс.)
1	5852
2	1199
3	406
4	201
5	119
6	80
7	55
8	41
9	35
10	25
Более 10	186

Блок-схема алгоритма представлена на рисунке 2. В первом блоке осуществляется сравнение библиографической записи каждого автора загружаемой статьи с известными системе написаниями фамилий и имен уже известных системе авторов. При поиске учитываются возможные ошибки, различные написания фамилий на разных языках, предыдущие фамилии автора и другие факторы. На основе результатов апробации первого варианта алгоритма, изложенного в работе [30], для быстрого поиска вариантов написания ФИО авторов в систему добавлена возможность кэширования таких вариантов альтернативных имен. Этот кэш пополняется автоматически по результатам разбора, но может быть отредактирован пользователем в своем личном кабинете. Результатам первого шага является список возможных IRID для каждого автора статьи. В этом же блоке осуществляется фильтрация по данным авторизации. Поскольку большинство статей вводится одним из соавторов, то при совпадении одного из возможных IRID с IRID пользователя, он считается однозначно определенным.

Далее производится циклическое повторение распознавания авторов, пока остается хотя бы один нераспознанный автор. Рассматривается два случая.

Если есть хотя бы один распознанный автор, то для всех возможных IRID нераспознанных авторов считается сумма ребер, связывающих эти IRID с распознанными авторами. Лучшим считается IRID с наибольшей суммой. Автор, соответствующий этому IRID, считается распознанным и все остальные IRID из его списка удаляются.

Возможно, что среди авторов на первом шаге нет ни одного распознанного. Для этого необходимо, чтобы статья вводилась в систему не одним из соавторов, и среди соавторов не было авторов с редкими фамилиями. Вследствие описанных выше причин вероятность такого события менее 7%. В этом случае производится более трудоемкое сравнение весов ребер всех возможных пар IRID всех соавторов. Пара с наибольшим ребром считается лучшей и для каждого из IRID этой пары автор считается распознанным и остальные IRID из его списка удаляются. Цикл повторяется, пока все авторы не будут распознаны.



Р и с. 2. Алгоритм распознавания IRID авторов

F i g. 2. IRID authors recognition algorithm

Результаты тестирования

Тестирование программной реализации алгоритма проводилось методом сравнения результатов автоматического расчета IRID и IRID, выбранных пользователями. Было обработано более 500 тысяч содержащихся в ИАС ИСТИНА публикаций. Результаты сравнения автоматического распознавания и ручного указания авторов представлены в таблице 2.

Таблица 2. **Сравнение автоматического и ручного определения авторов**

Table 2. **Comparison of automatic and manual identification of authors**

	Распознан	*Указан*	*Совпадение*	*Количество*
tp	+	+	+	85%
tn	-	-		3.2%
fp	+	+	-	3.3%
fn	-	+		0.5%
s	+	-		8%

Для 88.2% результаты автоматического расчета совпали с выбором пользователей, 3.8% результатов автоматического выбора IRID являются ошибочными с точки зрения пользователей, и для 8% случаев автоматический расчет рекомендовал IRID для авторов статей, которые не были указаны пользователем при вводе статьи. Таким образом, точность tp/(tp+fp) работы алгоритма составила 96%, полнота tp/(tp+fn) составила 99%, F-мера – 0.97. Отдельного рассмотрения требует категория s. Ручная проверка показывает, что в большинстве случаев автоматическое определение IRID сработало правильно, но при ручной обработке автор пользователем указан не был. В систему загружено большое количество статей, для которых из всех соавторов указан IRID только соавтора, осуществлявшего ввод статьи в систему. Это обстоятельство может быть обусловлено тем фактом, что пользователь при вводе статьи заинтересован в повышении только своего рейтинга и не хочет тратить время на ручной выбор всех соавторов. Использование автоматической рекомендации правильного IRID в таком случае значительно улучшит качество добавляемой в систему информации.

Заключение

Представленный в работе алгоритм позволяет производить анализ библиографического описания публикаций для идентификации авторов публикации и сопоставления каждому из авторов уникального идентификатора (IRID). Повышение полноты и точности определения IRID позволяет повысить качество получаемых наукометрических данных для принятия управленческих решений, а также уменьшить количество ошибок при определении прав доступа к данным на основе моделей.

Результаты работы алгоритма также могут использоваться для автоматического построения связей между объектами с целью описания правил доступа к данным в ABAC[21].

Список использованных источников

[1] Налимов, В. В. Наукометрия. Изучение науки как информационного процесса / В. В. Налимов, З. М. Мульченко. – М.: Наука, 1969. 192 с.

[2] Кириллова, О. В. Как оформить статью и научный журнал в целом для корректного индексирования в международных наукометрических базах данных / О. В. Кириллова. – DOI 10.24069/2542-0267-2018-1-2-52-72 // Научный редактор и издатель. – 2018. – Т. 3, № 1-2. – С. 52-72. – URL: https://www.scieditor.ru/jour/article/view/66/55 (дата обращения: 21.01.2020). – Рез. англ.

[3] Писляков, В. В. Зачем создавать национальные индексы цитирования? / В. В. Писляков // Научные и технические библиотеки. – 2007. – № 2. – С. 65-72. – URL: https://elibrary.ru/item.asp?id=9548771& (дата обращения: 21.01.2020).

[4] Еременко, Г. О. ELIBRARY.RU: курс на повышение качества контента / Г. О. Еременко // Университетская книга. – 2016. – № 3. – С. 62-68. – URL: https://elibrary.ru/item.asp?id=25721733 (дата обращения: 21.01.2020).

[5] Гуськов, А. Е. Российская наукометрия: обзор исследований / А. Е. Гуськов // Библиосфера. – 2015. – № 3. – С. 75-86. – URL: https://elibrary.ru/item.asp?id=24100709 (дата обращения: 21.01.2020). – Рез. англ.

[6] Мотрошилова, Н. В. Недоброкачественные сегменты наукометрии / Н. В. Мотрошилова // Вестник Российской академии наук. – 2011. – Т. 81, № 2. – С. 134-146. – URL: https://elibrary.ru/item.asp?id=16228786 (дата обращения: 21.01.2020). – Рез. англ.

[7] Богатов, В. В. Коллаборации научных организаций как элемент инфраструктуры науки / В. В. Богатов, Д. С. Сыроежкина // Наука. Инновации. Образование. – 2016. – Т. 11, № 4. – С. 30-44. – URL: https://elibrary.ru/item.asp?id=28123076 (дата обращения: 21.01.2020). – Рез. англ.

[8] Садовничий, В. А. Интеллектуальная система тематического исследования наукометрических данных: предпосылки создания и методология разработки. Часть 1 / В. А. Садовничий, В. А. Васенин. – DOI 10.17587/prin.9.51-58 // Программная инженерия. – 2018. – Т. 9, № 2. – С. 51-58. – URL: https://elibrary.ru/item.asp?id=32475947 (дата обращения: 21.01.2020). – Рез. англ.

[9] Васенин, В. А. Использование наукометрических информационно-аналитических систем для автоматизации проведения конкурсных процедур на примере информационно-аналитической системы «ИСТИНА» / В. А. Васенин, А. А. Зензинов, К. В. Лунев. – DOI 10.17587/prin.7.472-480 // Программная инженерия. – 2016. – Т. 7, № 10. – С. 472-480. – URL: https://elibrary.ru/item.asp?id=27179302 (дата обращения: 21.01.2020). – Рез. англ.

[10] Маршакова-Шайкевич, И. В. Тематический спектр исследовательской активности России / И. В. Маршакова-Шайкевич // Вестник Российской академии наук. – 2007. – Т. 77, № 9. – С. 811-818. – URL: https://elibrary.ru/item.asp?id=9552050 (дата обращения: 21.01.2020).

[11] Васенин, В. А. Интеллектуальная система тематического исследования наукометрических данных: состояние и перспективы / В. А. Васенин, С. А. Афонин, М. А. Зан-

чурин, А. А. Зензинов, А. С. Козицын, А. А. Коршунов, М. А. Кривчиков, Д. А. Шачнев // Материалы VII Международный конференции Знания – Онтологии – Теории (ЗОНТ-2019). – Новосибирск: ИМ СО РАН, 2019. – С. 94-103. – URL: https://elibrary.ru/item.asp?id=42432021 (дата обращения: 21.01.2020).

[12] A Unified Attribute-Based Access Control Model Covering DAC, MAC and RBAC / X. Jin, R. Krishnan, R. Sandhu. – DOI 10.1007/978-3-642-31540-4_4 // Data and Applications Security and Privacy XXVI. DBSec 2012. Lecture Notes in Computer Science; N. Cuppens-Boulahia, F. Cuppens, J. Garcia-Alfaro (ed.). Springer, Berlin, Heidelberg. – 2012. – Vol. 7371. – Pp. 41-55. – URL: https://link.springer.com/chapter/10.1007/978-3-642-31540-4_4 (дата обращения: 21.01.2020).

[13] Sandhu, R. S. Access control: principle and practice / R. S. Sandhu, P. Samarati. – DOI 10.1109/35.312842 // IEEE Communications Magazine. – 1994. – Vol. 32, No. 9. – Pp. 40-48. – URL: https://ieeexplore.ieee.org/document/312842 (дата обращения: 21.01.2020).

[14] Девянин, П. Н. Модели безопасности компьютерных систем / П. Н. Девянин. – М.: Академия, 2005. – 144 с.

[15] Гайдамакин, Н. А. Разграничение доступа к информации в компьютерных системах / Н. А. Гайдамакин. – Екатеринбург: Изд-во Урал. ун-та, 2003. – 328 с.

[16] Платонов, А. В. Методы автоматического построения онтологий / А.В. Платонов, Е.А. Полещук. – DOI 10.15827/0236-235X.114.047-052 // Программные продукты и системы. – 2016. – № 2. – С. 47-52. – URL: https://elibrary.ru/item.asp?id=28186543 (дата обращения: 21.01.2020). – Рез. англ.

[17] Бубарева, О.А. Исследование механизмов автоматического построения онтологий над множеством неструктурированных данных / О. А. Бубарева. – DOI 10.25699/SSSB.2019.25.27609 // Южно-Сибирский научный вестник. – 2019. – № 1. – С. 77-82. – URL: http://s-sibsb.ru/images/articles/2012/2/40_161- (дата обращения: 21.01.2020). – Рез. англ.

[18] Servos, D. Current Research and Open Problems in Attribute-Based Access Control / D. Servos, S.L. Osborn. – DOI: 10.1145/3007204 // ACM Computing Surveys. – 2017. – Vol. 49, No. 4. – Article No.: 65. – URL: https://dl.acm.org/doi/10.1145/3007204 (дата обращения: 21.01.2020).

[19] HGABAC: Towards a Formal Model of Hierarchical Attribute-Based Access Control / D. Servos, S. L. Osborn. – DOI 10.1007/978-3-319-17040-4_12 // Foundations and Practice of Security. FPS 2014. Lecture Notes in Computer Science; F. Cuppens, J. Garcia-Alfaro, N. Z. Heywood, P. W. L. Fong (ed.). Springer, Cham. – 2015. – Vol. 8930. – Pp. 187-204. – URL: https://link.springer.com/chapter/10.1007/978-3-319-17040-4_12 (дата обращения: 21.01.2020).

[20] Narouei, M. Automatic Extraction of Access Control Policies from Natural Language Documents / M. Narouei, H. Takabi, R. Nielsen. – DOI 10.1109/TDSC.2018.2818708 // IEEE Transactions on Dependable and Secure Computing. – 2020. – Vol. 17, issue 3. – Pp. 506-517. – URL: https://ieeexplore.ieee.org/document/8323229 (дата обращения: 21.01.2020).

[21] Afonin, S. Ontology Models for Access Control Systems / S. Afonin. – DOI 10.1109/RPC.2018.8482178 // 2018 3rd Russian-Pacific Conference on Computer Technology and Applications (RPC). – Vladivostok, 2018. – Pp. 1-6. – URL: https://ieeexplore.ieee.org/document/8482178 (дата обращения: 21.01.2020).

[22] Васенин, В. А. Внедрение реляционной модели логического разграничения доступа в web-приложения информационных систем, разработанных на основе библиотеки Django / В. А. Васенин, А.А. Иткес. – DOI 10.17587/prin.9.195-208 // Программная инженерия. – 2018. – Т. 9, № 5. – С. 195-208. – URL: https://elibrary.ru/item.asp?id=34994520 (дата обращения: 21.01.2020). – Рез. англ.

[23] Васенин, В. А. Архитектурно-технологические аспекты разработки и сопровождения больших информационно-аналитических систем в сфере науки и образования / В. А. Васенин, М. А. Занчурин, А. С. Козицын, М. А. Кривчиков, Д. А. Шачнев. – DOI 10.17587/prin.8.448-455 // Программная инженерия. – 2017. – Т. 8, № 10. – С. 448-455. – URL: https://elibrary.ru/item.asp?id=30397168 (дата обращения: 21.01.2020). – Рез. англ.

[24] Васенин, В. А. Интеллектуальная система тематического исследования научно-технической информации (ИСТИНА) / В. А. Васенин, С. А. Афонин, А. С. Козицын, Д. Д. Голомазов, А. В. Бахтин, Г. М. Ганкин // Обозрение прикладной и промышленной математики. – 2012. – Т. 19, № 2. – С. 239-240. – URL: https://istina.msu.ru/publications/article/813649 (дата обращения: 21.01.2020).

[25] Vasenin, V. Methods for Intelligent Data Analysis Based on Keywords and Implicit Relations: The Case of "ISTINA" Data Analysis System / V. Vasenin, K. Lunev, S. Afonin, D. Shachnev. – DOI 10.1109/APSSE47353.2019.00027 // 2019 Actual Problems of Systems and Software Engineering (APSSE). – Moscow, 2019. – Pp. 157-161. – URL: https://ieeexplore.ieee.org/document/8943809 (дата обращения: 21.01.2020).

[26] Determination of thematic proximity of scientific journals and conferences using big data technologies / A. Kozitsin, S. Afonin, D. Shachnev // Proceedings of the 21st Conference on Scientific Services & Internet (SSI-2019). CEUR Workshop Proceedings; M. Gorbunov-Posadov, A. Elizarov, M. Yakobovskiy (ed.). Novorossiysk-Abrau, Russia. – 2020. – Vol. 2543. – Pp. 407-413. – URL: http://ceur-ws.org/Vol-2543/spaper12.pdf (дата обращения: 21.01.2020).

[27] Васенин, В. А. Разрешение неоднозначности имен авторов: анализ публикаций / В. А. Васенин, А.Э. Гаспарянц. – DOI 10.17587/prin.8.264-275 // Программная инженерия. – 2017. – Т. 8, № 6. – С. 264-275. – URL: https://elibrary.ru/item.asp?id=29373361 (дата обращения: 21.01.2020). – Рез. англ.

[28] Афонин, С. А. Разрешение неоднозначности авторства публикаций при автоматической обработке библиографических данных / А. Э. Гаспарянц // Программная инженерия. – 2014. – № 1. – С. 25-28. – URL: https://elibrary.ru/item.asp?id=21431817 (дата обращения: 21.01.2020). – Рез. англ.

[29] Афонин, С. А. Автоматическое построение функции оценки качества в задаче разрешения неоднозначности имен авторов научных публикаций / А. Э. Гаспа-

рянц // Программная инженерия. – 2015. – № 10. – С. 31-37. – URL: https://elibrary.ru/item.asp?id=24365410 (дата обращения: 21.01.2020). – Рез. англ.

[30] Козицын, А. С. Разрешение неоднозначностей при определении авторов публикации с использование графов соавторства в больших коллекциях библиографических данных / А. С. Козицын, С. А. Афонин. – DOI 10.17587/prin.8.556-562 // Программная инженерия. – 2017. – Т. 8, № 12. – С. 556-562. – URL: https://elibrary.ru/item.asp?id=30710914 (дата обращения: 21.01.2020). – Рез. англ.

Поступила 21.01.2020; принята к публикации 20.04.2020; опубликована онлайн 25.05.2020.

Об авторах:

Козицын Александр Сергеевич, ведущий научный сотрудник Научно-исследовательского института механики, Московский государственный университет имени М. В. Ломоносова (119991, Россия, г. Москва, ГСП-1, Ленинские горы, д. 1), кандидат физико-математических наук, ORCID: http://orcid.org/0000-0002-8065-9061, alexanderkz@mail.ru

Афонин Сергей Александрович, ведущий научный сотрудник, Научно-исследовательского института механики, Московский государственный университет имени М. В. Ломоносова (119991, Россия, г. Москва, ГСП-1, Ленинские горы, д. 1), кандидат физико-математических наук, ORCID: http://orcid.org/0000-0003-3058-9269, serg@msu.ru

Все авторы прочитали и одобрили окончательный вариант рукописи.

References

[1] Nalimov V.V., Mulchenko Z.M. *Naukometriia. Izucheniia razvitiia nauki kak informatsionnogo protsessa* [Scientometrics. Study of the development of science as an information process]. Moscow, Nauka Publ.; 1969. (In Russ.)

[2] Kirillova O.V. How to arrange an article and scientific journal to avoid indexing errors in international scientometric databases. *Science Editor and Publisher*. 2018; 3(1-2):52-72. (In Russ., abstract in Eng.) DOI: https://doi.org/10.24069/2542-0267-2018-1-2-52-72

[3] Pislyakov V.V. Why Create National Citation Indexes? *Scientific and Technical Libraries*. 2007; (2):65-72. Available at: https://elibrary.ru/item.asp?id=9548771& (accessed 21.01.2020). (In Russ.)

[4] Eremenko G.O., ELIBRARY.RU: Course to Improve the Quality of Content. *Universitetskaia kniga*. 2016; (3):62-68. Available at: https://elibrary.ru/item.asp?id=25721733 (accessed 21.01.2020). (In Russ.)

[5] Guskov A.E. Russian Scientometrics: A Review of Researches. *Bibliosphere*. 2015; (3):75-86. Available at: https://elibrary.ru/item.asp?id=24100709 (accessed 21.01.2020). (In Russ., abstract in Eng.)

[6] Motroshilova N.V. Inferior Segments of the Scientometrics. *Vestnik Rossijskoj Akademii Nauk.* 2011; 81(2):134-146. Available at: https://elibrary.ru/item.asp?id=16228786 (accessed 21.01.2020). (In Russ., abstract in Eng.)

[7] Bogatov V.V., Syroezhkina D.S. Scientific Collaboration as an Element of Science Infrastructure. *Science. Innovation. Education*. 2016; 11(4):30-44. Available at: https://elibrary.ru/item.asp?id=28123076 (accessed 21.01.2020). (In Russ., abstract in Eng.)

[8] Sadovnichy V.A. Vasenin V.A. Intellectual System of Thematic Investigation of Scientometrical Data: Background of Creation and Methodology of Development. *Software Engineering*. 2018; 9(2):51-58. (In Russ., abstract in Eng.) DOI: https://doi.org/10.17587/prin.9.51-58

[9] Vasenin V.A., Zenzinov A.A., Lunev K.V. The Usage of CRIS-systems for the Contest Procedures Automation in Terms of the ISTINA Information System. *Software Engineering*. 2016; 7(10):472-480. (In Russ., abstract in Eng.) DOI: https://doi.org/10.17587/prin.7.472-480

[10] Marshakova-Shaikevich I.V. Thematic Spectrum of Research Activity in Russia. *Vestnik Rossijskoj Akademii Nauk.* 2007; 77(9):811-818. Available at: https://elibrary.ru/item.asp?id=9552050 (accessed 21.01.2020). (In Russ.)

[11] Vasenin V.A., Afonin S.A., Zanchurin M.A., Zenzinov A.A., Kozitsin A.S., Korshunov A.A., Krivchikov M.A., Shachnev D.A. Intellectual System of Thematic Investigation of Scientometrical Data: State and Prospects. In: Proceedings of the International conference "Knowledge – Ontology – Theories" (KONT-2019). IM SB RAS, Novosibirsk; 2019. p. 94-103. Available at: https://elibrary.ru/item.asp?id=42432021 (accessed 21.01.2020). (In Russ.)

[12] Jin X., Krishnan R., Sandhu R. A Unified Attribute-Based Access Control Model Covering DAC, MAC and RBAC. In: Cuppens-Boulahia N., Cuppens F., Garcia-Alfaro J. (ed.) Data and Applications Security and Privacy XXVI. DBSec 2012. *Lecture Notes in Computer Science*. 2012; 7371:41-55. Springer, Berlin. (In Eng.) DOI: https://doi.org/10.1007/978-3-642-31540-4_4

[13] Sandhu R.S., Samarati P. Access control: principle and practice. *IEEE Communications Magazine*. 1994; 32(9):40-48. (In Eng.) DOI: https://doi.org/10.1109/35.312842

[14] Devyanin P.N. *Modeli bezopasnosti komp'juternyh sistem* [Security model of computer systems]. Moscow: Publ. Center Academy; 2005. (In Russ.)

[15] Gaydamakin N.A. *Razgranichenie dostupa k informatsii v komp'yuternykh sistemakh* [Differentiation of Access to Information in Computer Systems]. Ekaterinburg, USU Publ.; 2003. (In Russ.)

[16] Platonov A.V., Poleschuk E.A. Methods of automatic ontology construction. *Software & Systems.* 2016; (2):47-52. (In Russ., abstract in Eng.) DOI: https://doi.org/10.15827/0236-235X.114.047-052

[17] Bubareva O.A. Reseach of Mechanisms of Automatic Construction of Ontologies over Multiple Unstructured Data. *South-Siberian Scientific Bulletin.* 2019; (1):77-82. (In Russ., abstract in Eng.) DOI: https://doi.org/10.25699/SSSB.2019.25.27609

[18] Servos D., Osborn S.L. Current Research and Open Problems in Attribute-Based Access Control. *ACM Computing Surveys*. 2017; 49(4):65. (In Eng.) DOI: https://doi.org/10.1145/3007204

[19] Servos D., Osborn S.L. HGABAC: Towards a Formal Model of Hierarchical Attribute-Based Access Control. In: Cuppens F., Garcia-Alfaro J., Heywood N.Z., Fong P.W.L. (ed.) Foun-

dations and Practice of Security. FPS 2014. *Lecture Notes in Computer Science*. 2015; 8930:187-204. Springer, Cham. (In Eng.) DOI: https://doi.org/10.1007/978-3-319-17040-4_12

[20] Narouei M., Takabi H., Nielsen R. Automatic Extraction of Access Control Policies from Natural Language Documents. *IEEE Transactions on Dependable and Secure Computing*. 2020; 17(3):506-517. (In Eng.) DOI: https://doi.org/10.1109/TDSC.2018.2818708

[21] Afonin S. Ontology Models for Access Control Systems. In: 2018 3rd Russian-Pacific Conference on Computer Technology and Applications (RPC), Vladivostok; 2018. pp. 1-6. (In Eng.) DOI: https://doi.org/10.1109/RPC.2018.8482178

[22] Vasenin V.A., Itkes A.A. Using Relation-Based Access Control Model within Django-Based Web Application. *Software Engineering*. 2018; 9(5):195-208. (In Russ., abstract in Eng.) DOI: https://doi.org/10.17587/prin.9.195-208

[23] Vasenin V.A., Zanchurin M.A., Kozitsin A.S., Krivchikov M.A., Shachnev D.A. Architectural and Technological Aspects of the Development and Maintenance of Large Information Analysis Systems in the Area of Science and Education. *Software Engineering*. 2017; 8(10):448-455. (In Russ., abstract in Eng.) DOI: https://doi.org/10.17587/prin.8.448-455

[24] Vasenin V.A., Afonin S.A., Kozitsin A.S., Golomazov D.D., Bahtin A.V., Gankin G.M. Intelligent System for Case Study of Scientific and Technical Information (ISTINA). *Obozrenie prikladnoi i promyshlennoi matematiki* = Review of Applied and Industrial Mathematics. 2012; 19(2):239-240. Available at: https://istina.msu.ru/publications/article/813649 (accessed 21.01.2020). (In Russ.)

[25] 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: 2019 Actual Problems of Systems and Software Engineering (APSSE), Moscow, Russia; 2019. p. 157-161. (In Eng.) DOI: https://doi.org/10.1109/APSSE47353.2019.00027

[26] Kozitsin A., Afonin S., Shachnev D. Determination of thematic proximity of scientific journals and conferences using big data technologies. In: Gorbunov-Posadov M., Elizarov A., Yakobovskiy M. (ed.) Proceedings of the 21st Conference on Scientific Services & Internet (SSI-2019). *CEUR Workshop Proceedings*. 2020; 2543:407-413. Novorossiysk-Abrau, Russia. Available at: http://ceur-ws.org/Vol-2543/spaper12.pdf (accessed 21.01.2020). (In Eng.)

[27] Vasenin V.A., Gaspariants A.E. Author Name Disambiguation: Analysis of Publications. *Software Engineering*. 2017; 8(6):264-275. (In Russ., abstract in Eng.) DOI: https://doi.org/10.17587/prin.8.264-275

[28] Afonin S.A., Gaspariants A.E. Scientific Article Authorship Disambiguation for Automated Bibliographic Records Processing. *Software Engineering*. 2014; (1):25-28. Available at: https://elibrary.ru/item.asp?id=21431817 (accessed 21.01.2020). (In Russ., abstract in Eng.)

[29] Afonin S.A., Gaspariants A.E. Construction of Quality Function for Scientific Papers Author Names Disambiguation Problem Using Supervised Learning Techniques. *Software Engineering*. 2015; (10):31-37. Available at: https://elibrary.ru/item.asp?id=24365410 (accessed 21.01.2020). (In Russ., abstract in Eng.)

[30] Kozitsin A.S., Afonin S.A. The Resolution of Ambiguities in the Identification of Authors of the Publication with the Use of Co-Authors' Graphs in Large Collections of Bibliographic Data. *Software Engineering*. 2017; 8(12):556-562. (In Russ., abstract in Eng.) DOI: https://doi.org/10.17587/prin.8.556-562

Submitted 21.01.2020; revised 20.04.2020; published online 25.05.2020.

About the authors:

Alexander S. Kozitsin, Leading Researcher of the Institute of Mechanics, Lomonosov Moscow State University (1, Leninskie gory, Moscow 119991, Russia), Ph.D. (Phys.-Math.), ORCID: http://orcid.org/0000-0002-8065-9061, alexanderkz@mail.ru

Sergey A. Afonin, Leading Researcher of the Institute of Mechanics, Lomonosov Moscow State University (1, Leninskie gory, Moscow 119991, Russia), Ph.D. (Phys.-Math.), ORCID: http://orcid.org/0000-0003-3058-9269, serg@msu.ru

All authors have read and approved the final manuscript.