УДК 004.912
DOI: 10.25559/SITITO.17.202101.711

Оригинальная статья

Алгоритм поиска по ключевым словам специалистов в заданной предметной области

А. С. Козицын*, С. А. Афонин, Д. А. Шачнев
ФГБОУ ВО «Московский государственный университет имени М. В. Ломоносова», г. Москва, Российская Федерация
119991, Российская Федерация, г. Москва, ГСП-1, Ленинские горы, д. 1
* alexanderkz@mail.ru

Аннотация

Особое внимание в современном мире уделяется тематическому поиску информации. Поиск специалистов можно назвать наиболее сложным направлением для тематического поиска, которое необходимо развивать для решения многих актуальных задач в сфере науки и бизнеса. В начале статьи дается краткий обзор существующих подходов к решению этой задачи, а также приводится список существующих информационных систем для поиска специалистов и экспертов. В работе описывается разработанный для решения этой задачи алгоритм, апробация которого проведена на данных наукометрической системы ИАС ИСТИНА, которая используется с 2012 года для сбора данных о научной деятельности в Московском государственном университете имени М. В. Ломоносова. Тематический поиск осуществляется на основании анализа данных о статьях, докладах и тезисах конференций, монографий, диссертаций, НИР, читаемых лекциях и спецкурсах. Для проведения тематического поиска специалистов используется следующая информация об объектах научной деятельности: название; список авторов; ключевые слова и аннотации. Основными особенностями разработанного алгоритма являются: возможность поиска специалистов без их активного участия в процессе поиска, возможность осуществлении поиска на русском и английском языках; использования для поиска не только научных публикаций, но и другой научной продукции; дополнительное определение авторитетности специалистов по графу соавторства. В статье приводится описание разработанного алгоритма тематического поиска, а также интерфейса его программной реализации, созданного в рамках апробации на реальных данных наукометрической системы ИСТИНА. В разработанном интерфейсе можно осуществлять поиск по ключевым словам с дополнительной фильтрацией по рубрикам, а также просматривать детализированную информацию по найденным специалистам.

Ключевые слова: наукометрия, информационные системы, тематический поиск, граф.

Авторы заявляют об отсутствии конфликта интересов.

Для цитирования: Козицын, А. С. Алгоритм поиска по ключевым словам специалистов в заданной предметной области / А. С. Козицын, С. А. Афонин, Д. А. Шачнев. – DOI 10.25559/SITITO.17.202101.711 // Современные информационные технологии и ИТ-образование. – 2021. – Т. 17, № 1. – С. 124-133.

Original article

Keyword Search Algorithm for Specialists in a Thematic Area

A. S. Kozitsin*, S. A. Afonin, D. A. Shachnev
Lomonosov Moscow State University, Moscow, Russian Federation
1 Leninskie gory, Moscow 119991, GSP-1, Russian Federation
*alexanderkz@mail.ru

Abstract

Thematic search for information is an important problem now. The search for specialists can be called the most difficult direction for thematic search. This direction needs to be developed in many areas of activity. At the beginning of the article, a brief overview of existing approaches to solving this problem is given. The following is a list of existing information systems for finding specialists and experts. The paper describes an algorithm developed for solving this problem. Approbation of the algorithm is implemented on the data of the scientometric system IAS ISTINA, which has been used since 2012 to collect data on scientific activities at the Lomonosov Moscow State University. Thematic search uses data analysis of articles, reports and abstracts of conferences, monographs, dissertations, lectures and training courses. The following information about the objects of scientific activity is used: name of object; list of authors; keywords and annotations. The main features of the developed algorithm are: the search for specialists without their active participation in the search process, the ability to search in Russian and English; use all kind of scientific products for searching; additional definition of the authority of specialists in the co-authorship graph. The article describes the developed thematic search algorithm and the interface of its software implementation. This interface was created to test the algorithm on real data of the scientometric system ISTINA. In this interface, you can search by keywords with additional filtering by headings and view detailed information on the specialists found.

Keywords: scientometrics, information systems, thematic search, graph.

The authors declare no conflict of interest.

For citation: Kozitsin A.S., Afonin S.A., Shachnev D.A. Keyword Search Algorithm for Specialists in a Thematic Area. *Sovremennye informacionnye tehnologii i IT-obrazovanie* = Modern Information Technologies and IT-Education. 2021; 17(1):124-133. DOI: https://doi.org/10.25559/SITITO.17.202101.711

Введение

В современном мире Интернет охватывает все области деятельности, и его можно называть отдельным информационным пространством [1]. В этом пространстве происходит интенсивное взаимодействие людей в науке, образовании, производстве и социальной сфере. Для обеспечения научной и преподавательской деятельности создаются специальные информационные системы, которые: позволяют экспертам реализовывать выполнение регламентов по проведению конкурсов в удаленном режиме [2]; обеспечивают студентов необходимой информации для успешного прохождения обучения, в том числе, в более предпочтительной для них электронной форме предоставления учебных материалов [3]; проводят анализ успешности обучения [4] и научной деятельности как на уровне организаций [5, 6], так и на уровне государства в целом [7]. Важным аспектом успешного функционирования таких систем является процесс автоматизации поиска информации [8] и обучение пользователей работе с информацией [9].

Особое внимание в современном мире уделяется тематическому поиску информации. Создаются специальные ресурсы, предоставляющие данные для обучения алгоритмов тематического поиска [10]. Разрабатываются алгоритмы для тематического поиска отдельных документов [11, 12] и библиографической информации [13], электронных ресурсов [14], журналов и конференций [15], элементов онтологий [16], а также авторитетных специалистов по заданным тематическим направлениям. Поиск специалистов можно назвать наиболее сложным направлением для тематического поиска, которое необходимо развивать для решения многих актуальных задач в сфере науки и бизнеса.

Поиск специалистов и консультантов по заданной тематике в наибольшей степени востребован среди представителей СМИ. Журналисты и владельцы информационных каналов ищут консультантов или авторов, для создания публикаций на заданную тему. В настоящее время существует множество различных информационных площадок, которые тем или иным образом позволяют решать такую задачу. Проект deadline.media позволяет зарегистрированным пользователям самостоятельно отмечать свои компетенции и указывать свои контакт для возможного поиска. Группы в социальных сетях, например, «Помогите журналисту» (www.facebook.com/groups/pomogitej) и «Герои и эксперты для СМИ» (www.facebook.com/groups/1792146334361293), позволяют самостоятельно искать специалистов среди участников групп и их знакомых. В работе [17] описывается способ автоматизированного сбора данных о человеке из социальных сетей.

На биржах статей, например, www.textsale.ru, Advego.ru, eTXT.ru, поиска квалифицированного специалиста можно разместить заказы на написание статей по заданной теме. Аналогичные биржи проектов, например, expertme.ru, позволяют размещать заявки на поиск экспертов для стартапов и дальнейшего развития коммерческих проектов.

Одним из примеров систем поиска специалистов в научной сфере является англоязычный проект network.expertisefinder.com, который осуществляет поиск по ключевым словам среди зарегистрированных сотрудников университетов США и Канады. Поиск производится по описаниям компетенций, которые представили сами сотрудники в описании своего профиля. Система автоматического сбора данных об экспертах реализована в китайском проекте ArnetMiner и Microsoft Academic Graph [18]. В этих проектах для поиска авторов используются ключевые слова статей из WoS.

В работе [19] описываются принципы формирования экспертных сетей, которые позволяют на основе взаимодействия отдельных узлов-участников формировать временные экспертные коллективы для решения тех или иных задач с перераспределением по экспертам потока вопросов в соответствии с их сложностью.

Русскоязычный проект РИНЦ [20] на основе библиографических данных позволяет искать специалистов по таким атрибутам профиля, как ФИО, страна, город, организация, уникальные коды различных систем (ORCID, Scopus, ResearcherId и другие) и по 70 тематическим рубрикам. Основным недостатком такого поиска является отсутствие возможности конкретизировать поиск по более детальным рубрикам или по ключевым словам.

Оценивая преимущества и недостатки описанных выше систем можно сформулировать требования, которым должна удовлетворять современная система тематического поиска специалистов в научной сфере. Система должна обрабатывать информацию на русском и на английском языках. Поиск должен производится по рубрикаторам для выбора глобальных тематических направлений, а также по ключевым для конкретизации области поиска[1] [21, 22, 23]. При показе результатов поиска необходимо осуществлять ранжирование с учетом соответствия условиям поиска и авторитетности специалистов. Система должна самостоятельно собирать данные о специалистах в научной сфере и не требовать обязательной самостоятельной регистрации. Основными источниками исходных данных для тематического анализа могут являться наукометрические системы [24], которые активно развиваются в настоящий момент в России [25]. Для решения этих задач авторами разработаны ряд методов и алгоритмов, которые апробированы в наукометрической системе ИСТИНА [26, 27].

Алгоритм поиска сотрудников

На странице поиска в ИАС ИСТИНА предусмотрен поиск сотрудников по ФИО, статей и журналов по названию, а также поиск высокотехнологичного оборудования и НИР. Такие возможности удобны при работе с системой, однако, не позволяют найти необходимую информацию, если известна только тематическая область поиска.

Разработанный алгоритм тематического поиска специалистов использует для поиска информацию о статьях, докладах и тезисах на конференциях, монографиях, НИР, читаемых лекции и спецкурсах, диссертациях и другие данные. Производится анализ названий объектов научно деятельности, списка соавторов, ключевых слов и аннотаций. Необходимые данные получаются автоматически из систем цитирования или вводятся в систему пользователями на этапе заполнения своих профи-

[1] *Бондарчук Д. В.* Алгоритмы интеллектуального поиска на основе метода категориальных векторов: дис. ... канд. физ.-мат. наук. Екатеринбург, 2016. 141 с.

лей. Алгоритм определения авторов по библиографическим описаниям объектов в процессе первичной загрузки данных в систему описан в работе [28]. Структура разработанного алгоритма поиска приведена на рис. 1.

На основе данных о загруженных в систему публикациях и других объектах научной продукции сотрудников строится граф соавторства. Его структура не зависит от поискового запроса, и процесс построения запускается периодически при появлении новых данных. Результатом работы этого шага алгоритма является граф с авторами-вершинами и взвешенными ребрами, вес которых определяется по частоте совместных публикаций каждой пары авторов.

Вторым шагом настройки алгоритма перед выполнением поисковых запросов является процесс определения степени соответствия авторов объектам поиска. В работе [29] показано, что для публикаций с большим количеством авторов для определения соответствия тематики статьи тематическим направлениям исследований авторов необходимо учитывать порядок авторов в библиографическом описании публикации. Аналогичный подход в разработанном алгоритме применяется и к другим объектам, в которых обнаружена такая закономерность. Результатом шага определения соответствия автора объектам поиска является двудольный взвешенный граф, связывающий авторов и объекты поиска. Такой граф также не зависит от поискового запроса и перестраивается по мере загрузки новых данных.

Р и с. 1. Схема алгоритма поиска сотрудников

F i g. 1. Employee search algorithm diagram

Этап поиска по ключевым словам является наиболее ресурсоемким с точки зрения объема обработки информации. Для быстрого выполнения запросов используется предварительная индексация названий объектов, их ключевых слов, текстов описаний и аннотаций с применением морфологического анализа. Результатом выполнения этого этапа обработки запроса является список объектов, содержащих слова запроса и число. Для каждого найденного объекта указывается его степень соответствия запросу. При этом учитывается не только частота встречаемости слов запроса в индексе объекта, но и позиция слов. Например, найденное слово в названии будет давать больший вес, чем слово, найденное в аннотации.

Для расширения области поиска возможно использование автоматического перевода ключевых слов, который применяется в других разработанных авторами алгоритмах, например, в алгоритме автоматического поиска переводов статей [30]. Однако, проведенные в рамках данного проекта эксперименты показали, что многозначность при переводе терминов значительно снижает качество поиска. При определении возможных правильных переводов в работе [30] сохранение точности обеспечивается за счет сверки возможных вариантов перевода с потенциальными правильными вариантами переводов названий статей, которую невозможно осуществить в рамках решения тематического поиска по словам.

Результаты поиска по ключевым словам могут проходить дополнительную фильтрацию по указанным пользователем рубрикам. Такая фильтрация помогает не только определить требуемы смысл многозначных слов, например, слово «граф», но и конкретизировать интересы пользователя при поиске по предметам исследования, которые изучаются в различных областях науки. Например, термин «сталь» может интересовать пользователя с точки зрения химии, механики или экономики, термин «информационная безопасность» может описывать как безопасность компьютерных систем, так и информационную безопасность подростков с точки зрения психологии. Дополнительное указание рубрики помогает предоставить пользователю более качественные результаты поиска.

На этапе ранжирование результатов поиска происходит вычисление веса соответствия каждого автора поисковому запросу. При построении веса предполагается, что хороший специалист в своей тематической области должен иметь объекты научной продукции разных типов. Например, если автор

имеет много докладов на конференциях по какой-либо теме, но не имеет научных статей и других объектов научной продукции, то его квалификация в данной области предполагается не очень высокой. В рамках такого подхода для подсчета суммарного ранга происходит предварительное ранжирование авторов по каждому типу научной продукции с учетом степени соответствия авторов объектам поиска и степени соответствия объектов запросу, полученных на втором и третьем шаге алгоритма. Результатом работы этого шага является ранжированный список авторов с указанием степени его соответствия поисковому запросу.

Для построения качественного рейтинга авторов и определения авторитетности необходимо учитывать не только их собственные работы, но и наличие у них взаимодействия с другими специалистами по исследуемой тематике, поскольку реализация серьезных научных исследований возможна только в составе научного коллектива с возможностью проведения совместных исследований, анализа и обсуждения результатов. Если автор в своих личных публикациях, докладах или лекциях часто упоминает определенный научный термин, но никто из его соавторов никогда не интересовался данной тематикой, то квалификация такого автора в этой научной области считается ниже, чем у членов больших научных коллективов с такой же степенью соответствия поисковому запросу по заданному термину. Уточнение рейтинга авторов производится с использованием подграфа соавторства, который строится из графа соавторства путем выделения подмножества вершин-авторов, имеющих ненулевую степень соответствия поисковому запросу. На основе полученного подграфа с использованием весов связей рассчитывается авторитетность авторов. Для расчета используется модификация алгоритма PageRank [31] в котором все ребра считаются двунаправленными.

Построение итогового рейтинга для каждого автора производится с использованием степени соответствия автора поисковому запросу и его авторитетности.

Апробация алгоритма

Для апробации описанного выше алгоритма была создана его программная реализация и интерфейс для анализа данных ИАС ИСТИНА, в котором пользователь может указать ключевые слова, рубрику и подразделение для поиска (Рис. 2).

Полученные данные представляются пользователю в виде отсортированного списка с указанием итогового рейтинга в каждой строке (Рис. 3).

Для получения более подробной информации о найденных сотрудниках пользователь может нажать на фамилию сотрудника для просмотра его полного профиля или нажать на значение рейтинга для просмотра списка научной продукции сотрудника (рис. 4), которая релевантна поисковому запросу, с возможностью последующего перехода на полное описание каждого элемента списка.

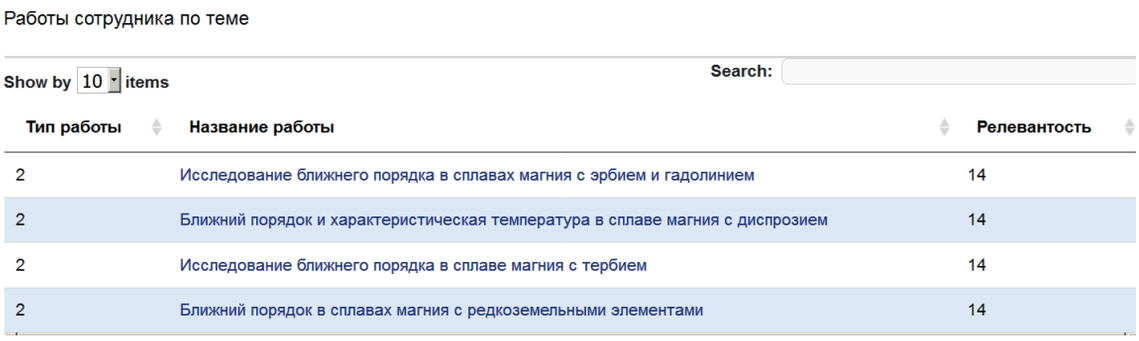
Р и с. 2. Интерфейс задания поискового запроса
F i g. 2. Search query interface

Список сотрудников по теме

Show by 10 items Search:

V	NAME	DEP
0.465	д.ф.-м.н. проф. Силонов Валентин Михайлович	Кафедра физики твердого тела(МГУ) Главный научный сотрудник
0.349	д.х.н. проф. Вербецкий Виктор Николаевич	Кафедра химической технологии и новых материалов(МГУ) Заведующий лабораторией
0.349	д.х.н. доц. Клямкин Семен Нисонович	Кафедра химической технологии и новых материалов(МГУ) Профессор
0.344	д.ф.-м.н. проф. Колобов Юрий Романович	Факультет фундаментальной физико-химической инженерии(Совм.)(МГУ) Профессор
0.236	к.х.н. Тарасов Борис Петрович	None

Р и с. 3. Представление результатов поиска
F i g. 3. Presentation of search results



Работы сотрудника по теме

Show by 10 items

Search:

Тип работы	Название работы	Релевантость
2	Исследование ближнего порядка в сплавах магния с эрбием и гадолинием	14
2	Ближний порядок и характеристическая температура в сплаве магния с диспрозием	14
2	Исследование ближнего порядка в сплаве магния с тербием	14
2	Ближний порядок в сплавах магния с редкоземельными элементами	14

Р и с. 4. Список работ сотрудника по теме запроса

F i g. 4. List of employee's works on the subject of the request

При тестировании результатов запроса «информационная безопасность» в рубрике «Информатика» для оценки использовался список top20 из найденных 50 сотрудников. В результатах запросе специалистами в указанной области оказались 17 человек (85%). Для запроса «титан» без указания рубрики точность оказалась значительно ниже (60%) из-за наличия у физиков работ с большим количеством соавторов 8 человек и более.

Представленный функционал могут использовать научные сотрудники для поиска дополнительной информации по своей тематике исследований или по смежным областям знаний, студенты при выборе кафедры и поиска научного руководителя, а также при подборе обучающих материалов, представители бизнеса при поиске потенциальных исполнителей для наукоемких проектов. Возможности такого поиска в комплексе с другими возможностями наукометрических систем помогают более качественно представить результаты деятельности организации в Интернет.

Заключение

Представленный в работе алгоритм позволяет производить тематический поиск специалистов по данным, содержащимся в наукометрических системах. Преимуществом алгоритма является возможность поиска специалистов без их активного участия в процессе поиска, возможность осуществлении поиска на русском и английском языках, использования для поиска не только научных публикаций, но и другой научной продукции, а также дополнительное определение авторитетности специалистов по графу соавторства.

Список использованных источников

[1] О едином цифровом пространстве научных знаний / А. Б. Антопольский, Н. Е. Каленов, В. А. Серебряков, А. Н. Сотников. – DOI 10.31857/S0869-5873897728-735 // Вестник Российской академии наук. – 2019. – Т. 89, № 7. – С. 728-735. – Рез. англ.

[2] Власова, С. А. Web-ориентированная система формирования контента единого цифрового пространства научных знаний / С. А. Власова, Н. Е. Каленов, А. Н. Сотников. – DOI 10.15827/0236-235X.131.365-374 // Программные продукты и системы. – 2020. – Т. 33, № 3. – С. 365-374. – Рез. англ.

[3] Голубева, Е. А. Использование потенциала электронных библиотечных систем в образовательной деятельности вуза / Е. А. Голубева, М. В. Смагина. – DOI 10.317773/2078-1768-2020-50-211-218 // Вестник Кемеровского государственного университета культуры и искусств. – 2020. – № 50. – С. 211-218. – Рез. англ.

[4] Asif, R. Analyzing undergraduate students' performance using educational data mining / R. Asif [и др.]. – DOI 10.1016/j.compedu.2017.05.007 // Computers & Education. – 2017. – Vol. 113. – Pp. 177-194.

[5] 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. – Рез. англ.

[6] 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. – Рез. англ.

[7] 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: 09.02.2021). – Рез. англ.

[8] Афонин, С. А. Использование систем семантического анализа для организации поиска научно-технической информации / С. А. Афонин, Д. Д. Голомазов, А. С. Козицын // Программная инженерия. – 2012. – № 2. – С. 29-34. – URL: https://elibrary.ru/item.asp?id=17588117 (дата обращения: 09.02.2021). – Рез. англ.

[9] Bradley, P. Expert Internet Searching / P. Bradley. – 5th Edition. – Facet Publishing, 2017.

[10] Сикулер, Д. В. Поиск данных для апробации интеллектуальных алгоритмов и технологий / Д. В. Сикулер // Символ науки. – 2020. – № 4. – С. 49-54. – URL: https://elibrary.ru/item.asp?id=42726648 (дата обращения: 09.02.2021).

[11] Mohammad, M. Automatic knowledge extraction from OCR documents using hierarchical document analysis / M. Mohammad, S. Kosaraju, T. Bayramoglu, G. Modgil, M. Kang. – DOI 10.1145/3264746.3264793 // Proceedings of the 2018 Conference on Research in Adaptive and Convergent Systems (RACS'18). – Association for Computing Machinery, New York, NY, USA, 2018. – Pp. 189-194.

[12] Chen, G. Ensemble application of convolutional and recurrent neural networks for multi-label text categorization / G. Chen, D. Ye, Z. Xing, J. Chen, E. Cambria. – DOI 10.1109/IJCNN.2017.7966144 // 2017 International Joint Conference on Neural Networks (IJCNN). – Anchorage, AK, USA: IEEE, 2017. – Pp. 2377-2383.

[13] Глазкова, А. В. Сравнение нейросетевых моделей для классификации текстовых фрагментов, содержащих биографическую информацию / А. В. Глазкова. – DOI 10.15827/0236-235X.126.263-267 // Программные продукты и системы. – 2019. – Т. 32, № 2. – С. 263-267. – Рез. англ.

[14] Гринченков, Д. В. Архитектура системы тематического поиска электронных образовательных ресурсов в сети интернет / Д. В. Гринченков, Д. Н. Кущий // Информатизация и связь. – 2016. – № 3. – С. 209-212. – URL: https://elibrary.ru/item.asp?id=26623194 (дата обращения: 09.02.2021). – Рез. англ.

[15] Kozitsin, A. Determination of thematic proximity of scientific journals and conferences using big data technologies / A. Kozitsin, S. Afonin, D. Shachnev // CEUR Workshop Proceedings. – 2020. – Vol. 2543. – Pp. 407-413. – URL: http://ceur-ws.org/Vol-2543/spaper12.pdf (дата обращения: 09.02.2021).

[16] Афонин, С. А. Программные механизмы агрегации данных, основанные на онтологическом представлении структуры реляционной базы наукометрических данных / С. А. Афонин, А. С. Козицын, Д. А. Шачнев. – DOI 10.17587/prin.7.408-413 // Программная инженерия. – 2016. – Т. 7, № 9. – С. 408-413. – Рез. англ.

[17] Самойлов, А. Ю. Разработка средств автоматизации поиска информации о человеке в открытых источниках сети Интернет / А. Ю. Самойлов, Е. Ю. Никитина. – DOI 10.17072/1993-0550-2020-1-74-79 // Вестник Пермского университета. Серия: Математика. Механика. Информатика. – 2020. – № 1(48). – С. 74-79. – Рез. англ.

[18] Berger, M. Effective Distributed Representations for Academic Expert Search / M. Berger, J. Zavrel, P. Groth. – DOI 10.18653/v1/2020.sdp-1.7 // Proceedings of the First Workshop on Scholarly Document Processing. – Association for Computational Linguistics, 2020. – Pp. 56-71.

[19] Ma, L. Performance Bounds of Decentralized Search in Expert Networks for Query Answering / L. Ma, M. Srivatsa, D. Cansever, X. Yan, S. Kase, M. Vanni. – DOI 10.1145/3300230 // ACM Transactions on Knowledge Discovery from Data. – 2019. – Vol. 13, No. 2. – Article 18.

[20] 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: 09.02.2021).

[21] 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 2019). – Moscow, IEEE, 2019. – Pp. 157-161.

[22] Шундеев, А. С. Об изменении размерности векторного представления текстовых данных / А. С. Шундеев. – DOI 10.17587/prin.10.265-273 // Программная инженерия. – 2019. – Т. 10, № 6. – С. 265-273. – Рез. англ.

[23] Wu, S. High dimensional data Clustering Algorithm Based on Sparse Feature Vector for Categorical Attributes / S. Wu, G. Wei. – DOI 10.1109/ICLSIM.2010.5461099 // 2010 International Conference on Logistics Systems and Intelligent Management (ICLSIM). – Harbin, China: IEEE, 2010. – Pp. 973-976.

[24] 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.

[25] 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: 09.02.2021). – Рез. англ.

[26] Интеллектуальная система тематического исследования наукометрических данных: состояние и перспективы / В. А. Васенин, С. А. Афонин, М. А. Занчурин, 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: 09.02.2021).

[27] 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: 09.02.2021).

[28] Козицын, А. С. Разрешение неоднозначностей при определении авторов публикации с использованием 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. – Рез. англ.

[29] 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. – Рез. англ.

[30] Козицын, А. С. Алгоритм определения переводов статей с использованием статистических данных / А. С. Козицын, С. А. Афонин, А. А. Зензинов // Электронные библиотеки. – 2018. – Т. 21, № 6. – С. 494-505. – URL: https://elibrary.ru/item.asp?id=37028485 (дата обращения: 09.02.2021). – Рез. англ.

[31] Gleich, D. F. PageRank Beyond the Web / D. F. Gleich. – DOI 10.1137/140976649 // SIAM Review. – 2015. – Vol. 57, issue 3. – Pp. 321-363.

Поступила 09.02.2021; одобрена после рецензирования 26.03.2021; принята к публикации 02.04.2021.

Об авторах:

Козицын Александр Сергеевич, ведущий научный сотрудник Научно-исследовательского института механики, ФГБОУ ВО «Московский государственный университет имени М. В. Ломоносова» (119991, Российская Федерация, 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, ФГБОУ ВО «Московский государственный университет имени М. В. Ломоносова» (119991, Российская Федерация, г. Москва, ГСП-1, Ленинские горы, д. 1), кандидат физико-математических наук, **ORCID:** http://orcid.org/0000-0003-3058-9269, serg@msu.ru

Шачнев Дмитрий Алексеевич, программист Научно-исследовательского института механики, ФГБОУ ВО «Московский государственный университет имени М. В. Ломоносова» (119991, Российская Федерация, г. Москва, ГСП-1, Ленинские горы, д. 1), **ORCID:** http://orcid.org/0000-0002-5940-9180, mitya57@gmail.com

Все авторы прочитали и одобрили окончательный вариант рукописи.

References

[1] Antopolskii A.B., Kalenov N.E., Serebryakov V.A., Sotnikov A.N. *O edinom tsifrovom prostranstve nauchnykh znanii* [Common Digital Space of Scientific Knowledge]. *Vestnik Rossijskoj akademii nauk*. 2019; 89(7):728-735. (In Russ., abstract in Eng.) DOI: https://doi.org/10.31857/S0869-5873897728-735

[2] Vlasova S.A., Kalenov N.E., Sotnikov A.N. *Web-orientirovannaia sistema formirovaniia kontenta edinogo tsifrovogo prostranstva nauchnykh znanii* [A web-based content generation system for a common digital space of scientific knowledge]. *Programmnye produkty i sistemy* = Software & Systems. 2020; 33(3):365-374. (In Russ., abstract in Eng.) DOI: https://doi.org/10.15827/0236-235X.131.365-374

[3] Golubeva E.A., Smagina M.V. *Ispolzovanie potentsiala elektronnykh bibliotechnykh sistem v obrazovatelnoi deiatelnosti vuza* [Using the potential of electronic library systems in the educational activities of universities]. *Bulletin of Kemerovo State University of Culture and Arts*. 2020; (50):211-218. (In Russ., abstract in Eng.) DOI: https://doi.org/10.317773/2078-1768-2020-50-211-218

[4] Asif R. et al. Analyzing undergraduate students' performance using educational data mining. *Computers & Education*. 2017; 113:177-194. (In Eng.) DOI: https://doi.org/10.1016/j.compedu.2017.05.007

[5] Vasenin V.A., Zenzinov A.A., Lunev K.V. *Ispolzovanie naukometricheskikh informatsionno-analiticheskikh sistem dlia avtomatizatsii provedeniia konkursnykh protsedur na primere informatsionno-analiticheskoi sistemy "ISTINA"* [The Usage of CRIS-systems for the Contest Procedures Automation in Terms of the ISTINA Information System]. *Programmnaia inzheneriia* = Software Engineering. 2016; 7(10):472-480. (In Russ., abstract in Eng.) DOI: https://doi.org/10.17587/prin.7.472-480

[6] Vasenin V.A., Zanchurin M.A., Kozitsin A.S., Krivchikov M.A., Shachnev D.A. *Arkhitekturno-tekhnologicheskie aspekty razrabotki i soprovozhdeniia bolshikh informatsionno-analiticheskikh sistem v sfere nauki i obrazovaniia* [Architectural and Technological Aspects of the Development and Maintenance of Large Information Analysis Systems in the Area of Science and Education]. *Programmnaia inzheneriia* = Software Engineering. 2017; 8(10):448-455. (In Russ., abstract in Eng.) DOI: https://doi.org/10.17587/prin.8.448-455

[7] Marshakova-Shaikevich I.V. *Tematicheskii spektr issledovatelskoi aktivnosti Rossii // Vestnik Rossiiskoi akademii nauk* [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 09.02.2021). (In Russ.)

[8] Afonin S.A., Golomazov D.D., Kozitsyn A.S., *Ispolzovanie sistem semanticheskogo analiza dlia orgmanizatsii poiska nauchno-tekhnicheskoi informatsii* [Using analytical ontology-based information systems for scientific and technological information search]. *Programmnaia inzheneriia* = Software Engineering. 2012; (2):29-34. Available at: https://elibrary.ru/item.asp?id=17588117 (accessed 09.02.2021). (In Russ., abstract in Eng.)

[9] Bradley P. Expert Internet Searching. 5th Ed. Facet Publishing; 2017. (In Eng.)

[10] Sikuler D.V. *Poisk dannykh dlia aprobatsii intellektualnykh algoritmov i tekhnologii* [Data search for testing intelligent algorithms and technologies]. *Simvol nauki* = Symbol of Science. 2020; (4):49-54. Available at: https://elibrary.ru/item.asp?id=42726648 (accessed 09.02.2021). (In Russ.)

[11] Mohammad M., Kosaraju S., Bayramoglu T., Modgil G., Kang M. Automatic knowledge extraction from OCR documents using hierarchical document analysis. *Proceedings of the 2018 Conference on Research in Adaptive and Convergent Systems (RACS'18).* Association for Computing Machinery, New York, NY, USA; 2018. p. 189-194. (In Eng.) DOI: https://doi.org/10.1145/3264746.3264793

[12] Chen G., Ye D., Xing Z., Chen J., Cambria E. Ensemble application of convolutional and recurrent neural networks for multi-label text categorization. *2017 International Joint Conference on Neural Networks (IJCNN).* IEEE, Anchorage, AK, USA; 2017. p. 2377-2383. (In Eng.) DOI: https://doi.org/10.1109/IJCNN.2017.7966144

[13] Glazkova A.V. *Sravnenie neirosetevykh modelei dlia klassifikatsii tekstovykh fragmentov, soderzhashchikh biograficheskuiu informatsiiu* [Comparison of neural network models for classifying text fragments containing biographical information]. *Programmnye produkty i sistemy* = Software & Systems. 2019; 32(2):263-267. (In Russ., abstract in Eng.) DOI: https://doi.org/10.15827/0236-235X.126.263-267

[14] Grinchenkov D.V., Kushchiy D.N. *Arkhitektura sistemy tematicheskogo poiska elektronnykh obrazovatelnykh resursov v seti Internet* [Architecture of system the subject search of electronic educational resources on the Internet]. *Informatizatsiia i sviaz* = Informatization and communication. 2016; (3):209-212. Available at: https://elibrary.ru/item.asp?id=26623194 (accessed 09.02.2021). (In Russ., abstract in Eng.)

[15] Kozitsin A., Afonin S., Shachnev D. Determination of thematic proximity of scientific journals and conferences using big data technologies. *CEUR Workshop Proceedings*. 2020; 2543:407-413. Available at: http://ceur-ws.org/Vol-2543/spa-per12.pdf (accessed 09.02.2021). (In Eng.)

[16] Afonin S.A., Kozitsyn A.S., Shachnev D.A. *Programmnye mekhanizmy agregatsii dannykh, osnovannye na ontologicheskom predstavlenii struktury reliatsionnoi bazy naukometricheskikh dannykh* [Software mechanisms for scientometrical data aggregation based on ontological representation of the relational database structure]. *Programmnaia inzheneriia* = Software Engineering. 2016; 7(9):408-413. (In Russ., abstract in Eng.) DOI: https://doi.org/10.17587/prin.7.408-413

[17] Samoilov A.Yu., Nikitina E.Yu. *Razrabotka sredstv avtomatizatsii poiska informatsii o cheloveke v otkrytykh istochnikakh seti internet* [Development of tools for automated search of information about a person in open web sources]. *Vestnik Permskogo universiteta Matematika. Mekhanika* = Bulletin of Perm University. Mathematics. Mechanics. Computer Science. 2020; (1):74-79. (In Russ., abstract in Eng.) DOI: https://doi.org/10.17072/1993-0550-2020-1-74-79

[18] Berger M., Zavrel J., Groth P. Effective Distributed Representations for Academic Expert Search. *Proceedings of the First Workshop on Scholarly Document Processing*. Association for Computational Linguistics; 2020. p. 56-71. (In Eng.) DOI: https://doi.org/10.18653/v1/2020.sdp-1.7

[19] Ma L., Srivatsa M., Cansever D., Yan X., Kase S., Vanni M. Performance Bounds of Decentralized Search in Expert Networks for Query Answering. *ACM Transactions on Knowledge Discovery from Data*. 2019; 13(2):18. (In Eng.) DOI: https://doi.org/10.1145/3300230

[20] Eremenko G.O. *ELIBRARY.RU: kurs na povyshenie kachestva kontenta* [ELIBRARY.RU: Course to Improve the Quality of Content]. *Universitetskaia kniga* = University Book. 2016; (3):62-68. Available at: https://elibrary.ru/item.asp?id=25721733 (accessed 09.02.2021). (In Russ.)

[21] 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. *2019 Actual Problems of Systems and Software Engineering (APSSE 2019).* IEEE, Moscow, Russia; 2019. p. 157-161. (In Eng.) DOI: https://doi.org/10.1109/APSSE47353.2019.00027

[22] Shundeev A.S. *Ob izmenenii razmernosti vektornogo predstavleniia tekstovykh dannykh* [On changing the dimension of the document embeddings]. *Programmnaia inzheneriia* = Software Engineering. 2019; 10(6):265-273. (In Russ., abstract in Eng.) DOI: https://doi.org/10.17587/prin.10.265-273

[23] Wu S., Wei G. High dimensional data Clustering Algorithm Based on Sparse Feature Vector for Categorical Attributes. *2010 International Conference on Logistics Systems and Intelligent Management (ICLSIM).* IEEE, Harbin, China; 2010. p. 973-976. (In Eng.) DOI: https://doi.org/10.1109/ICLSIM.2010.5461099

[24] Sadovnichy V.A. Vasenin V.A. *Intellektualnaia sistema tematicheskogo issledovaniia naukometricheskikh dannykh: predposylki sozdaniia i metodologiia razrabotki. Chast 1* [Intellectual System of Thematic Investigation of Scientometrical Data: Background of Creation and Methodology of Development]. *Programmnaia inzheneriia* = Software Engineering. 2018; 9(2):51-58. (In Russ., abstract in Eng.) DOI: https://doi.org/10.17587/prin.9.51-58

[25] Guskov A. E., *Rossiiskaia naukometriia: obzor issledovanii* [Russian Scientometrics: A Review of Research-es]. *Bibliosfera* = Bibliosphere. 2015; (3):75-86. Available at: https://elibrary.ru/item.asp?id=24100709 (accessed 09.02.2021). (In Russ., abstract in Eng.)

[26] Vasenin V.A., Afonin S.A., Zanchurin M.A., Zenzinov A.A., Kozitsin A.S., Korshunov A.A., Krivchikov M.A., Shachnev D.A. *Intellektualnaia sistema tematicheskogo issledovaniia naukometricheskikh dannykh: sostoianie i perspektivy* [Intellectual System of Thematic Investigation of Scientometrical Data: State and Prospects]. *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 09.02.2021). (In Russ.)

[27] Vasenin V.A., Afonin S.A., Kozitsin A.S., Golomazov D.D., Bahtin A.V., Gankin G.M. *Intellektualnaia sistema tematicheskogo issledovaniia nauchno-tekhnicheskoi informatsii (ISTINA*) [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 09.02.2021). (In Russ.)

[28] Kozitsin A.S., Afonin S.A. *Razreshenie neodnoznachnostei pri opredelenii avtorov publikatsii s ispolzovaniem grafov soavtorstva v bolshikh kollektsiiakh bibliograficheskikh dannykh* [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]. *Programmnaia inzheneriia* = Software Engineering. 2017; 8(12):556-562. (In Russ., abstract in Eng.) DOI: https://doi.org/10.17587/prin.8.556-562

[29] Kozitsin A.S., Afonin S.A. *Algoritm razresheniia neodnoznachnosti imen avtorov v IAS ISTINA* [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. (In Russ., abstract in Eng.) DOI: https://doi.org/10.25559/SITITO.16.202001.108-117

[30] Kozitsin A.S., Afonin S.A., Zenzinov A.A. *Algoritm opredeleniia perevodov statei s ispolzovaniem statisticheskikh dannykh* [Algorithm for linking translated articles using authorship statistics]. *Elektronnye biblioteki* = Russian Digital Libraries Journal. 2018; 21(6):494-505. Available at: https://elibrary.ru/item.asp?id=37028485 (accessed 09.02.2021). (In Russ., abstract in Eng.)

[31] Gleich D.F. PageRank Beyond the Web. *SIAM Review*. 2015; 57(3):321-363. (In Eng.) DOI: https://doi.org/10.1137/140976649

Submitted 09.02.2021; approved after reviewing 26.03.2021; accepted for publication 02.04.2021.

About the authors:

Alexander S. Kozitsin, Leading Researcher of the Institute of Mechanics, Lomonosov Moscow State University (1 Leninskie gory, Moscow 119991, GSP-1, Russian Federation), 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, GSP-1, Russian Federation), Ph.D. (Phys.-Math.), **ORCID:** http://orcid.org/0000-0003-3058-9269, serg@msu.ru

Dmitry A. Shachnev, Software Developer of the Institute of Mechanics, Lomonosov Moscow State University (1 Leninskie gory, Moscow 119991, GSP-1, Russian Federation), **ORCID:** http://orcid.org/0000-0002-5940-9180, mitya57@gmail.com

All authors have read and approved the final manuscript.