УДК 004.912

ОПРЕДЕЛЕНИЕ ТЕМАТИЧЕСКОЙ БЛИЗОСТИ НАУЧНЫХ ЖУРНАЛОВ И КОНФЕРЕНЦИЙ С ИСПОЛЬЗОВАНИЕМ АНАЛИЗА ГРАФА СОАВТОРСТВА

А. С. Козицын, С. А. Афонин, Д. А. Шачнев

НИИ механики МГУ им. М.В. Ломоносова, г. Москва

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

Аннотация

Количество публикуемых в мире журналов очень велико. В этой связи, необходим программный инструментарий, который позволит анализировать тематические связи журналов. Разработанный авторами и представленный в этой работе алгоритм 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 разработанном для этих целей интерфейсе 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*.

1. МОДЕЛЬ

Количество публикуемых в настоящий момент научных журналов очень велико. Например, в информационно-аналитической системе (ИАС) «ИСТИНА» [1] зарегистрировано более 70 тысяч научных журналов и еще более 200 тысяч различных сборников научных публикаций и материалов конференций. В этой связи молодым ученым, аспирантам и студентам необходимы сервисы, которые позволят автоматически подбирать журналы, которые наиболее соответствуют по тематике их научным интересам. Для решения этой задачи может использоваться аккумулированный опыт всего научного сообщества.

Существует несколько возможных способов решения задачи определения тематической близости журналов, позволяющих в автоматическом режиме определить меру сходства между каждой парой журналов. Первый способ основан на использовании тематического анализа полнотекстовых данных, таких, как полнотекстовые описания журналов, тексты опубликованных в журналах статей, их полных или кратких аннотаций, а также указанных авторами ключевых слов. Для проведения тематического анализ полнотекстовых данных в настоящее время разработан широкий спектр различных методов: использование деревьев решений [2]; преобразование текста документа в вектор в многомерном пространстве [3] с использованием частотных характеристик слов [4] и последующим применением геометрических методов классификации (SVN, K-means и аналогичные); нейронные сети. Следует отметить, что в большинстве методов полнотекстовой классификации требуется проведение предобработки текстов, в том числе с использованием методов морфологического анализа [5], которые существенно зависят от языка текста, и требуют предварительной настройки на каждый из используемых языков.

На основе результатов проведенного полнотекстового тематического анализа с использованием указанных методов классификации или кластеризации полных текстов публикаций или их аннотаций возможно построение оценки смысловой близости публикаций в журнале с описанной информационной потребностью конкретного пользователя. Область своих научных интересов пользователь может описать при помощи задания достаточного количества ключевых

слов или загрузить в систему полные тексты своих статей для автоматического построения тематического портрета пользователя. В рамках такого подхода для проведения анализа необходимо иметь достаточно точно описанные тематические профили всех журналов или полные тексты статей, публикуемых в этих журналах.

Получение достаточно полных полнотекстовых данных является сложной задачей, поскольку во многих журналах открытая публикация полных текстов статей не разрешена. Вместе с тем, использование только ключевых слов для проведения тематического анализа может давать слишком общие результаты. В первую очередь это объясняется тем фактом, что во многих случаях ключевые слова статьи характеризуют в большей степени не ее тематику, а связь статьи с одним из приоритетных направлений развития науки, технологий и техники в Российской Федерации. Например, ключевое слово «Нанотехнология», которое часто упоминается в приоритетных направлениях развития науки, технологий и техники в РФ, встречается в статьях совершенно различной тематики: «Разработка новой медицинской нанотехнологии для поражения раковых клеток при детских острых лимфобластных лейкозах»; «Разработка и производство новых наноструктурированных алмазоподобных углеродных покрытий трибологического назначения»; «Разработка и создание сверхчувствительных полевых и зарядовых наноструктур для считывающих и сенсорных устройств наноэлектроники»; «Использование радионуклидов и источников ионизирующего излучения в нанохимии, ядерной медицине и для исследования процессов, происходящих в окружающей среде». Таким образом, тематическая классификация статей по данному ключевому слову определит не столько тематику статьи, сколько участие авторов статьи в проектах по определенному приоритетному направлению. В этой связи использование полнотекстового тематического анализа для решения поставленной выше задачи в наукометрических системах может сталкиваться с определенными трудностями.

Альтернативным методом оценки тематической близости журналов является анализ графа соавторства статей, публикуемых в этих журналах. Такой подход может использоваться как автономно, так и совместно с методами анализа по ключевым словам [6] или текстам. Граф соавторства – это двудольный граф, в котором множество вершин-авторов связано ребрами с множеством вершин-статей. При реализации такого подхода предполагается, что в большинстве случаев

авторы публикуют свои результаты проводимых научных исследований в тематически близких журналах. Вследствие этого в близких по тематике журналах большое количество статей связано в графе соавторов путем длины 2. Основанный на использовании графов соавторства подход, в отличие от методов полнотекстового тематического анализа, не требует наличия полнотекстовой информации о статьях и использует только библиографические данные статей, публикуемых в журналах. Такие данные могут быть получены из наукометрических систем (например, ИАС «ИСТИНА») или систем цитирования (например, WoS).

2. АЛГОРИТМ ОЦЕНКИ ТЕМАТИЧЕСКОЙ БЛИЗОСТИ ЖУРНАЛОВ

Формально задачу оценки близости журналов можно сформулировать следующим образом. Необходимо построить граф, вершинами которого являются журналы, а веса ребер соответствуют их тематической близости.

Разработанный алгоритм на первом шаге для каждой пары журналов вычисляет все пары статей, опубликованных в этих журналах одним автором. Если паре журналов соответствует только одна пара статей, то такие пары считаются не связанными. Если паре журналов соответствует несколько пар статей, то журналы считаются связанными ребром с определенным весом.

В рамках настоящей работы рассматривалось несколько методов определения веса ребра. Наиболее простым методом является определение веса ребра равным количеству уникальных авторов среди соответствующих пар статей. Основным недостатком такого метода является невозможность учитывать значимость авторов для каждой статьи. Во многих случаях статьи пишутся только одним автором, фамилия которого ставится на первом месте в ее библиографическом описании. Остальные соавторы могут участвовать в работе над статьей незначительно, и их основное направление научной деятельности может не совпадать с ее тематикой.

Для проверки гипотезы о значимости порядка авторов при проведении тематического анализа была проведена оценка доли статей, в которых порядок авторов определяется лексикографическим порядком, а не значимостью в работе над статьей. Из наукометрической системы МГУ были отобраны для анализа все статьи в журналах за 2014–2017 гг. с количеством авторов от 2 до 7.

Для каждого из указанного количества авторов были посчитаны проценты

статей L, для которых правильный набор авторов определяется лексикографическим порядком. Результаты расчета для различного количества авторов приведены в таблице 1. Из данных, приведенных в таблице, можно сделать вывод, что в большинстве случаев основным автором является тот, который указан в библиографическом описании первым. Для учета этого факта была разработана формула расчета веса ребер с учетом позиции автора в библиографическом описании статьи. Вес автора для каждой статьи определяется как 1/2+1/(2K) для первого автора и 1/(2K) для остальных соаторов, где K – количество соавторов в статье. Степень связи по заданному автору для двух журналов определяется как минимум из максимумов его весов по подмножествам статей в каждом из журналов. Окончательный вес ребра связи между двумя журналами может быть рассчитан как сумма степеней их связи по всем авторам.

Количество авторов	L
2	24%
3	16%
4	9%
5	6%
6	6%
7	3%

Таблица 1. Процент статей с лексикографическим порядком.

3. ПРОГРАММНАЯ РЕАЛИЗАЦИЯ И РЕЗУЛЬТАТЫ ТЕСТИРОВАНИЯ

При выборе языка для программной реализации алгоритма учитывались такие особенности алгоритма, как большой объем обрабатываемых данных, необходимость быстрого доступа к хранящимся в СУБД данным, небольшие требования к объемам памяти для создания временных структур данных и отсутствие необходимости вести диалог с пользователем. Учитывая эти требования, для реализации был выбран язык PL/SQL. Расчет тематической близости между журналами производится с заданными интервалами времени и сохраняется в таблицы СУБД.

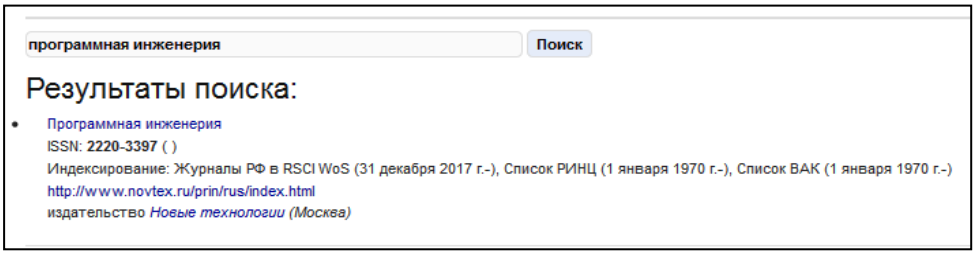

Рис. 1. Интерфейс контекстного поиска журналов

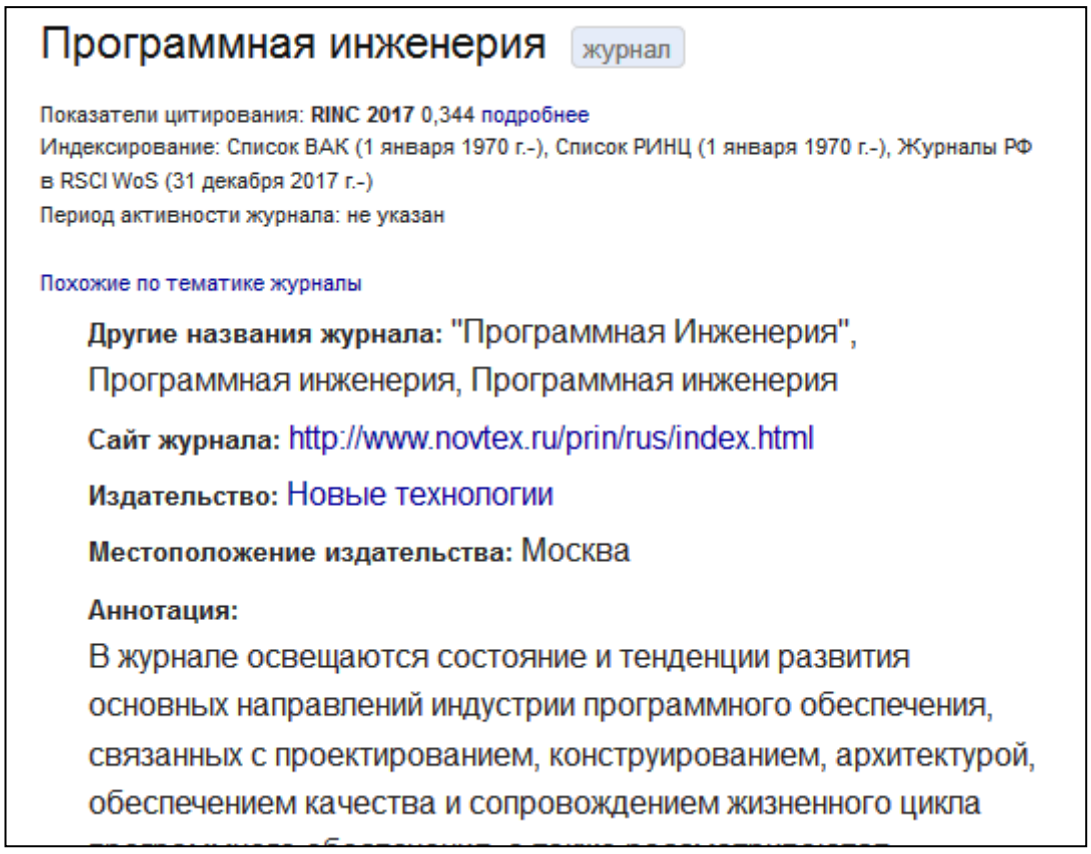

Рис. 2. Карточка журнала

В разработанном для этих целей интерфейсе [7] пользователь может выбрать один журнал, близкий ему по тематике (рис. 1, 2), 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 (рис. 3). Веб-интерфейс реализован с использованием открытой библиотеки DataTables [8]. В информационную карточку каждого журнала добавлена ссылка для перехода к таблице со списком тематически похожих журналов. В этой таблице указываются названия близких по тематике журналов и меры сходства. Кроме того, для возможности быстрой оценки

авторитетности каждого журнала из списка в таблице приводятся данные о количестве публикаций в этом журнале за 5 лет (зарегистрированных в системе «ИСТИНА»), а также данные Web of Science и РИНЦ. С целью удобной навигации по графу близости журналов в разработанном интерфейсе также реализована возможность перехода по ссылкам на список похожих журналов непосредственно из каждого элемента списка. Средствами библиотеки DataTables для быстрого поиска по названиям журналов реализован механизм быстрой фильтрации по части названия журнала.

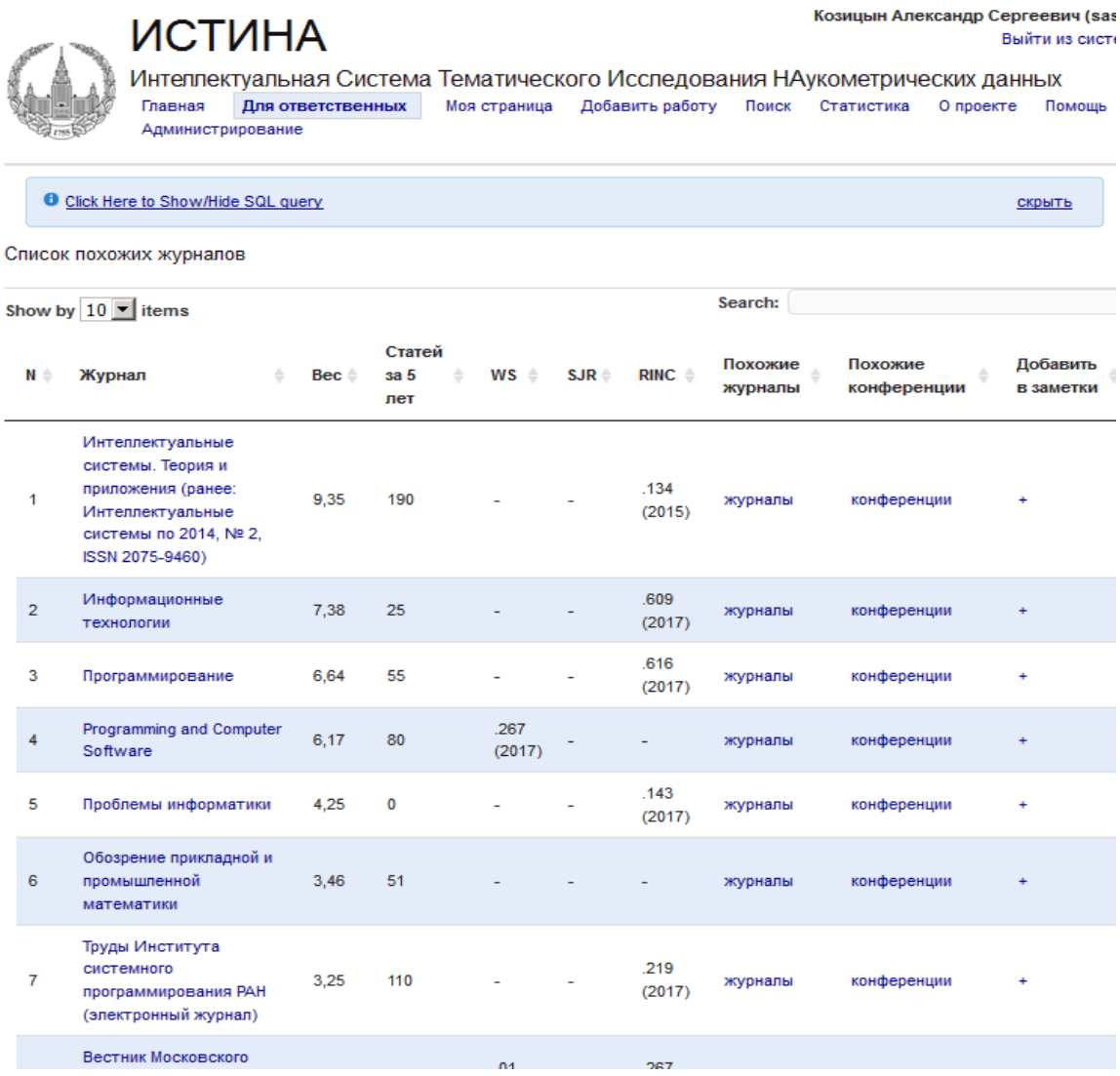

Рис. 3. Поиск похожих по тематике журналов

Для удобства работы пользователя предусмотрена возможность добавлять выбранный журнал в заметки, которые впоследствии можно просматривать, редактировать, а также использовать при последующем поиске. Дополнительно предоставляется возможность подбора похожих по тематике конференций (рис.

4).

Список похожих на журнал конференций

Show by 10 items Search:

N	Конференция	Вес	Количество докладов	Похожие конференции	Похожие журналы
1	Ломоносовские чтения - 2018. Секция "Механика"(2018)	2,54	129	конференции	журналы
2	Ломоносовские чтения-2018, секция "Вычислительная математика и кибернетика"(2018)	1,42	134	конференции	журналы
3	Знания-Онтологии-Теории (ЗОНТ-2017)(2017)	1,17	5	конференции	журналы
4	Знания-Онтологии-Теории (ЗОНТ-2019)(2019)	1,13	3	конференции	журналы
5	«Ломоносовские чтения - 2019». Секция «ВМК»(2019)	1,04	111	конференции	журналы
6	Научный сервис в сети Интернет 2019(2019)	0,95	5	конференции	журналы
7	«Modern Network Technologies, MoNeTec-2018»:(2018)	0,92	7	конференции	журналы
8	2018 Annual International Conference on Biologically Inspired Cognitive Architectures Ninth Annual Meeting of the BICA Society, Ninth Annual Meeting of the BICA Society(2018)	0,84	6	конференции	журналы

Рис. 4. Поиск похожих по тематике конференций

Тестирование разработанной программной реализации алгоритма проводилось по следующей методике. Из полученных результатов случайным образом было выбрано 200 пар связей журналов. Экспертами была проведена ручная оценка совпадения тематик журналов с простановкой баллов (2 – точная; 1 – не совсем точная; 0 – ошибочная). Общая сумма баллов делилась на удвоенное количество анализируемых связей. Оценка точности по этой методике составила 78%.

В качестве примера ошибок алгоритма можно привести, например, список журналов, которые определены как близкие по тематике к изданию «Труды Высшей школы Министерства внутренних дел СССР»: «Философские науки»; «Логические исследования»; «Известия МГТУ МАМИ»; «Логико-философские исследования»; «Вестник Московского университета. Серия 7: Философия». Такие ошибки могут возникать как следствие слишком широкой тематической области принимаемых в журнал статей.

ЗАКЛЮЧЕНИЕ

Алгоритм, описанный в настоящей работе, позволяет автоматически оценивать степень тематической близости научных журналов на основе библиографического описания статей и без использования полнотекстовых версий статей. Следует отметить, что алгоритм нечувствителен к языку журнала и подбирает похожие журналы на разных языках, что сложно реализуемо для алгоритмов, основанных на анализе полнотекстовой информации.

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 [9].

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

СПИСОК ЛИТЕРАТУРЫ

1. *Садовничий В.А., Васенин В.А.* Интеллектуальная система тематического исследования наукометрических данных: предпосылки создания и методология разработки. Часть 1 // Программная инженерия. 2018. Т. 9. № 2. С. 51–58.

2. Воронцов К.В. Лекции по логическим алгоритмам классификации URL:http://www.ccas.ru/voron/download/LogicAlgs.pdf

3. Шундеев А. С. Об изменении размерности векторного представления текстовых данных. Программная инженерия, , 2019 Т. 10. № 6. С. 265-273.

4. *Бурлаева Е.И. Павлыш В.Н.* Анализ методов преобразования текстов в форму объектов векторного пространства//Программная инженерия. 2019, Т. 10. № 1. с.30-37.

5. *Трофимов И.В.* Морфологический анализ русского языка: обзор прикладного характера//Программная инженерия. 2019, Т. 10. № 9. с. 391–399

6. 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 Actual Problems of Systems and Software Engineering – APSSE 2019, IEEE Conference Proceedings, pages 151-155, United States, 2019

7. ИАС ИСТИНА. URL: https://istina.msu.ru.
8. Библиотека datatables. URL: https://datatables.net/
9. *Afonin S.* Ontology models for access control systems. In 2018 3rd Russian-Pacific Conference on Computer Technology and Applications (RPC). 2018. P. 1–6. doi: 10.1109/RPC.2018.8482178.

DETERMINING THE THEMATIC PROXIMITY OF SCIENTIFIC JOURNALS AND CONFERENCES USING BIG DATA TECHNOLOGIES

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

Institute of Mechcanics Lomonosov Moscow State University, Moscow

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

Abstract

The number of scientific journals published in the world is very large. In this regard, it is necessary to create software tools that will allow analyzing thematic links of journals. The algorithm presented in this paper uses graphs of co-authorship for analyzing the thematic proximity of journals. It is insensitive to the language of the journal and can find similar journals in different languages. This task is difficult for algorithms based on the analysis of full-text information. Approbation of the algorithm was carried out in the scientometric system IAS ISTINA. Using a special interface, a user can select one interesting journal. Then the system will automatically generate a selection of journals that may be of interest to the user. In the future, the developed algorithm can be adapted to search for similar conferences, collections of publications and research projects. The use of such tools will increase the publication activity of young employees, increase the citation of articles and quoting between journals. In addition, the results of the algorithm for determining thematic proximity between journals, collections, conferences and research projects can be used to build rules in the ontology models for access control systems.

Keywords*: thematic classification, bibliographic data, graph of co-authorship, Information Systems*

REFERENCES

1. *Sadovnichii V.A., Vasenin V.A.* Intellektualnaia sistema tematicheskogo issledovaniia naukometricheskikh dannykh: predposylki sozdaniia i metodologiia razrabotki. Chast 1 // Programmnaia inzheneriia. 2018. T. 9. No 2. P. 51–58.

2. Voroncov K.V. Lekcii po logicheskim algoritmam klassifikacii. URL:http://www.ccas.ru/voron/download/LogicAlgs.pdf.

3. Shundeev A.S., Ob izmenenii razmernosti vectornogo predstavlenija tekstovyh dannyh// Programmnaia inzheneriia. 2019, T.10. No 6. p.265-273.

4. Burlaeva E.I., Pavlysh V.N., Analiz metodov preobrazovaniya tekstov v formu obyektov vektornogo prostarnstva// Programmnaia inzheneriia. 2019. T. 10. No 1. S. 30–37.

5. Trofimov I.V. Morfologicheskiy analiz russkogo yazyka: obzor prikladnogo haraktera// Programmnaia inzheneriia. 2019. T. 10. No 9. S. 391–399.

6. Vasenin V., Lunev K., Afonin S., Shachnev D. Methods for intelligent data aalysis based on keywords and implicit relations: The case of "istina" data analysis system//In Actual Problems of Systems and Software Engineering - APSSE 2019, IEEE Conference Proceedings, pages 151–155, United States, 2019.

7. IAS ISTINA. URL: https://istina.msu.ru.

8. Datatables. URL: https://datatables.net/

9. *Afonin S.* Ontology models for access control systems. In 2018 3rd Russian-Pacific Conference on Computer Technology and Applications (RPC). 2018. P. 1–6. doi: 10.1109/RPC.2018.8482178

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 KOZITSIN – 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.

Dmitiy 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 12 ноября 2019 года