УДК 004.4

АЛГОРИТМ ОПРЕДЕЛЕНИЯ ПЕРЕВОДОВ СТАТЕЙ С ИСПОЛЬЗОВАНИЕМ СТАТИСТИЧЕСКИХ ДАННЫХ

А. С. Козицын[1], С. А. Афонин[2], А. А. Зензинов[3]

НИИ механики Московского государственного университета им. М.В. Ломоносова, 119991, Российская Федерация, Москва, Ленинские горы, д. 1

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

Аннотация

В настоящее время происходит активное внедрение наукометрических систем для автоматизации процесса анализа эффективности деятельности научных организаций с целью применения различных методов стимулирования научной деятельности. Одними из наиболее важных индикаторов являются количество публикаций и их цитируемость. Для оценки этих показателей необходимы средства автоматизированного построения связей между оригинальными статьями и их переводами. В настоящей работе проанализированы существующие методы оценки близости оригинального текста и его возможного перевода, показана их недостаточная эффективность для построения связей между статьями и описаня разработанный авторами метод автоматического поиска переводов статей в больших коллекциях библиографических данных. Особенностью разработанного алгоритма является использование статистических данных о публикации статей в различных журналах и информации о соавторах анализируемых статей. Представленный в настоящей работе алгоритм позволяет осуществлять поиск переводов статей без предварительной настройки на заданные пары языков оригинала и перевода статьи, а также не требует использования больших коллекций обучающих выборок. Апробация программной реализации алгоритма проводилась в наукометрической системе Московского государственного университета (МГУ) им. М.В. Ломоносова. Результаты тестирования показали ее достаточную эффективность и возможность использования разработанного алгоритма для автоматического построения рекомендаций пользователям для отметки в системе переводных версий статей.

***Ключевые слова**: библиографические данные, анализ графов, перевод, статья, статистика, наукометрия, цитирование, автоматизированные системы.*

ВВЕДЕНИЕ

В современном мире наблюдается устойчивая тенденция развития средств автоматизации оценки деятельности персонала во всех сферах экономической деятельности. Основной целью такой автоматизации является создание более эффективных механизмов поощрения и стимулирования исполнителей при решении поставленных перед ними задач. Список показателей, которые учитываются для проведения подобных оценок, существенно зависят от анализируемой области деятельности. В методике определения показателей эффективности деятельности в научной сфере (наукометрии) [1] используются многочисленные измерения и рассчитанные по ним статистические данные, учитывающие количество и качество производимой сотрудниками научной продукции: научных публикаций, докладов на конференциях, патентов и других показателей.

Учет количества подобной продукции в рамках небольшой научной организации (до 50 человек) может проводиться путем сбора бумажных отчетов о работе сотрудников или использованием простых средств автоматизации, например, таблиц в Excel.

Для больших организаций необходимо использование систем автоматизации сбора и анализа показателей научной информации. В этой области существует значительное количество небольших систем, позволяющих осуществлять централизованный сбор и предварительную обработку таких показателей [2–5], а также создаются системы корпоративного уровня, которые позволяют производить не только сбор и агрегацию информации, но и ее использование на всех уровнях управления организацией [6]. Такие системы имеют сложную архитектуру [7], включают в себя механизмы интеллектуального анализа и средства обеспечения безопасности доступа к данным [8, 9] и используют большой набор различных наукометрических показателей и индексов для оценки результатов научной деятельности.

КОЛИЧЕСТВО ПУБЛИКАЦИЙ И ПОКАЗАТЕЛИ ЦИТИРОВНИЯ

Традиционно одним из основных количественных показателей результатов научной деятельности являются количество публикаций и различные индексы цитирования, отражающие качество и авторитетность публикаций в научном сообществе. К основным индексам цитирования можно отнести следующие [10].

Количество ссылок на статью. Показатель рассчитывается по одной из баз данных Web of Science; Scopus; Google Scholar или РИНЦ. Его значение может существенно отличаться в зависимости от выбранной базы. Например, при расчете количества ссылок по РИНЦ будут учитываться ссылки из русскоязычных изданий, в то время как Web of Science больше ориентирован на англоязычные издания. Google Scholar учитывает ссылки из статей, которые размещены в интернете, но могут быть не опубликованы в журналах, и т. д. Также в различных метриках при расчете количества ссылок могут учитываться или не учитываться ссылки самоцитирования или цитирования соавторов, использоваться разные периоды времени для учета ссылок, учитываться области размещения ссылок и быть другие отличия.

Импакт-фактор. Показатель рассчитывается для журнала и показывает авторитетность издания, в котором опубликована статья. В отличие от количества ссылок на статью, этот показатель не оценивает востребованность самой статьи. Качество статьи оценивается в предположении, что авторитетные издания не будут публиковать статьи без серьезного рецензирования. Основным преимуществом этого показателя по сравнению с предыдущим является возможность оценки статьи непосредственно после публикации. Значение импакт-фактора рассчитывается на основе двух показателей: количества статей в журнале за предыдущие несколько лет и количества ссылок на эти статьи, сделанные в текущем году. Отношение этих величин можно назвать «средней цитируемостью» статей в журнале. Как правило, при оценках научных работ используется импакт-фактор за трехлетний период, рассчитанный по базе Web of Science.

Индекс Хирша. Показатель рассчитывается для автора и совмещает показатели количества статей и количества цитирований на каждую из статей. Например, значение индекса Хирша 6 означает, что у автора имеется 6 статей, которые цитируются не менее 6 раз. Значения данного показателя для одного автора также могут значительно отличаться в зависимости от базы, по которой производится расчет

количества ссылок (Web of Science, Scopus, Google Scholar или РИНЦ), учета самоцитирования и других факторов.

Для вычисления всех перечисленных выше наукометрических показателей используются число статей и число цитирований на статью. Точное определение этих параметров играет существенную роль при построении объективных наукометрических оценок. Вместе с тем, многие журналы печатают переводные версии статей, которые не является самостоятельной статьей и не должны учитываться как отдельная статья, однако ссылки, сделанные на переводную версию статьи, должны увеличивать наукометрические показатели ее авторов. В некоторых случаях такие переводы осуществляются без участия автора статьи. В этой связи возникает задача автоматизации процесса сопоставления оригинальной версии статьи и всех ее переводов.

ОПРЕДЕЛЕНИЕ ПЕРЕВОДОВ СТАТЕЙ

Сложность определения переводных версий статей обусловлена тем, что предоставление информации о статье и ее переводе может осуществляться не только в разное время по мере выхода изданий, но и получаться из разных источников. Некоторые журналы на своих страницах в интернете размещают информацию о наличии у них переводных изданий, однако такая информация плохо структурирована, а ее автоматическая обработка очень сложна. Кроме того, значительная часть переводов размещается в иностранных изданиях, не являющихся полными переводными версиями соответствующих русскоязычных изданий.

Задача автоматического перевода названий статей является очень трудоемкой, поскольку в названиях используются многозначные слова, и необходимо при переводе учитывать специфику предметной области статьи. В качестве примера можно привести результаты автоматического перевода названия статьи двумя популярными переводчиками. Оригинал статьи – «Степень самоочищения агродерново-подзолистых супесчаных почв, удобренных осадком сточных вод», переводная версия статьи – «Self-Purification of Agrosoddy-Podzolic Sandy Loamy Soils Fertilized with Sewage Sludge», обратный перевод Гугл [11] – «Самоочищение Агрозодди-Подзолик Сэнди глинистые почвы, оплодотворенные с отстоем сточных вод». Оригинал статьи – «Поверхностные волны Релея и Лява при отрицательном коэффициенте Пуассона изотропных сред», переводная версия статьи – «Rayleigh and Love surface

waves in isotropic media with negative Poisson's ratio», обратный перевод Промт [12] – «Рэлей и Любовные волны поверхности в изотропических СМИ с отношением отрицательного Пуассона».

Следует отметить, что в большинстве случаев имеется смысловое сходство автоматического перевода и перевода, сделанного автором, но набор слов существенно различается. Это объясняется, в первую очередь, неоднозначностью терминов в любом языке. В одних случаях в языке перевода отсутствуют полностью эквивалентные термины языка оригинала, в других – автоматическая система выбирает не совсем верные термины.

В настоящее время, в связи усилением борьбы с плагиатом, активно развивается направление поиска эквивалентных текстов на разных языках, обсуждаемых, в том числе, на конференции «Обнаружение заимствований» [13]. Например, в системе «Антиплагиат» создан модуль «Переводные заимствования», который способен определять степень эквивалентности текстов, написанных на разных языках. Используемый в системе метод анализа основан на понятии *n*-грамм. Элементами *n*-грамм являются классы эквивалентных слов, что позволяет учитывать наличие эквивалентных терминов в разных языках [14]. Такой подход эффективен для поиска переводов полных текстов, но имеет ряд существенных недостатков, которые затрудняют его использование для поиска переводных статей по названиям.

Во-первых, построение классов эквивалентных слов требует настройки под каждую пару языков. В системе «Антиплагиат» используется только русско-английский перевод, а в случае перевода статей необходимо учитывать все возможные языки.

Во-вторых, использование *n*-грамм возможно только для достаточно длинных частей текста и плохо применимо к названиям статей.

Альтернативным подходом к автоматизации процесса поиска переводных версий статей является использование статистических данных о распределении статей по журналам и информации об авторах статей. Такой подход позволяет находить возможные переводы статей, основываясь только на структуре связей в графе соавторства статей, без использования информации о языке оригинала и перевода статьи. Результаты поиска могут вручную подтверждаться экспертом на основе сравнения названий статей. Использование графа соавторства успешно применяется в задачах идентификации авторов [15]. Расширение этого подхода и использование

существующих связей между статьями и журналами позволяет выявлять дополнительные закономерности в данных с целью выявления скрытых связей между объектами. Ниже представлено описание алгоритма, который определяет связи между переводными и оригинальными версиями статей на основе анализа графа соавторства.

Основой разработанного алгоритма является предположение, что оригинальная статья и ее перевод должны быть опубликованы одним и тем же авторским коллективом с разницей не более года в журналах на разных языках.

После построения пар статей, которые могут являться переводами, производится построение двудольного графа журналов, которые печатают переводные статьи. Метрика для оценки степени связи журналов в графе строится на основе мощности множеств статей в каждом из журналов и мощности множества пар статей в этих журналах, которые могут являться переводами. Результатом работы этого этапа алгоритма является множество пар журналов, в одном из которых часто печатаются переводные статьи из второго журнала. В процессе работы граф связей журналов уточняется на основе вносимых пользователями данных о своих статьях. Для этого используется как явное указание пользователями связей между оригинальной и переводной статьями, так и информация о DOI статьи, задаваемых авторами. Многие авторы указывают библиографические данные оригинала статьи в русскоязычном журнале, внося индекс DOI переводной версии для учета ссылок из Web of Science. Таким образом, собрав из внешних источников информацию о статье, по DOI можно точно определить название переводного журнала для указанного в статье русскоязычного журнала.

На основе построенного множества журналов производится поиск возможного перевода статьи. Поиск осуществляется среди статей, которые могут являться переводами (имеют совпадающее множество авторов, а дата публикации отличается не более чем на год) и опубликованы в журналах, связанных ребром в построенном ранее графе журналов.

Для апробации алгоритма использовались данные о публикациях сотрудников МГУ им. М.В. Ломоносова. Авторами статьи разработан модуль, добавленный в функционал наукометрической системы организации [16]. Разработанный для этих целей интерфейс (рис. 1) позволяет экспертам проводить оценку результатов

работы модуля и отмечать в системе правильные и ошибочные варианты предлагаемых переводов.

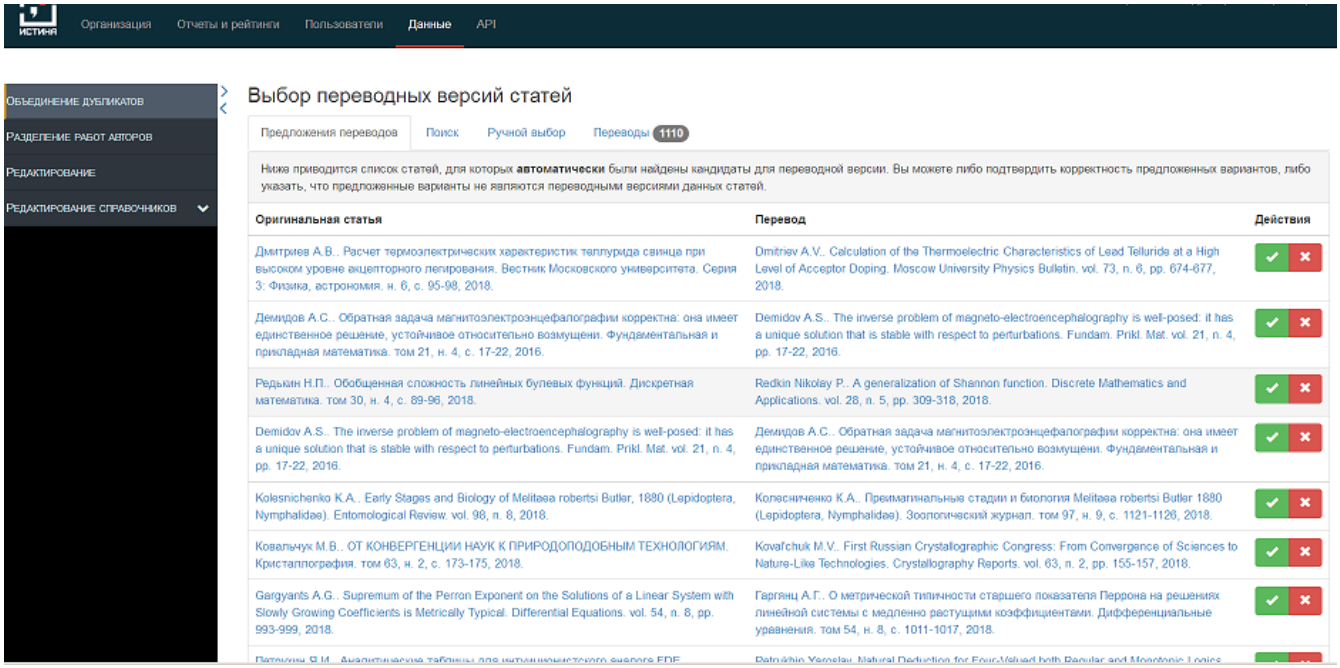


Рис. 1. Модуль выбора переводных версий статей

Для удобства работы предусмотрена возможность проведения поиска по статье, заданной пользователем (рис. 2).

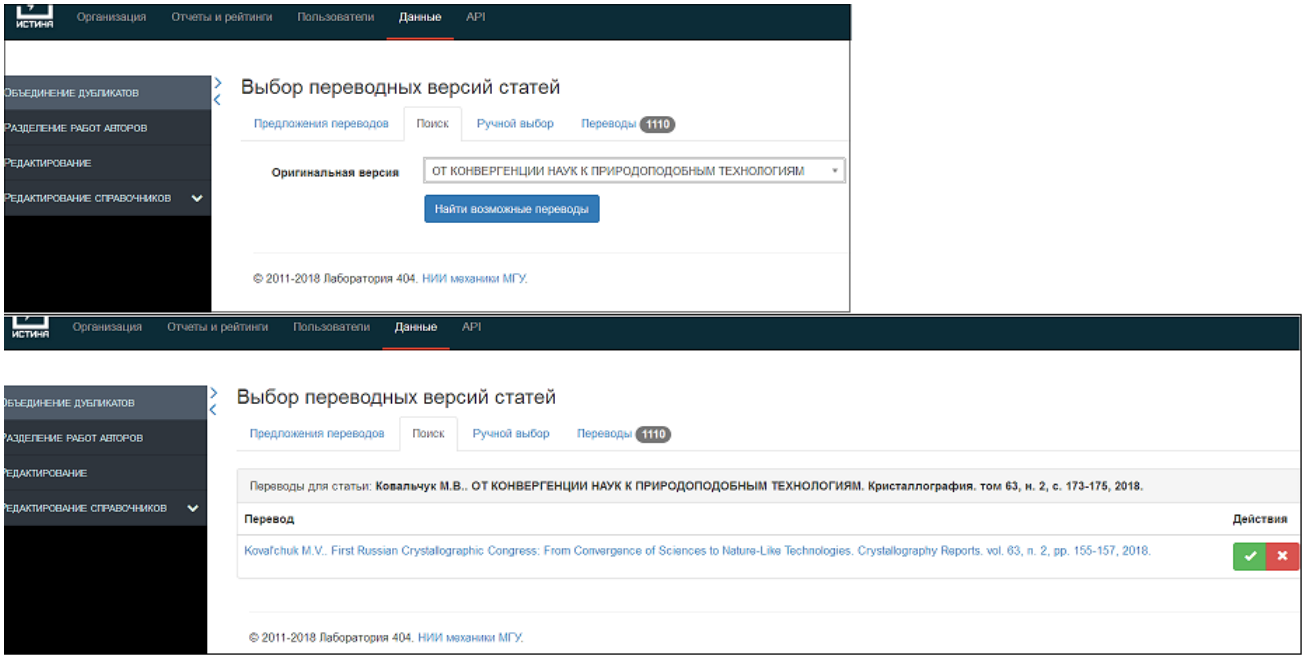


Рис. 2. Поиск по статье, заданной пользователем

Следует отметить, что алгоритм может использоваться как для обработки полной коллекции статей, так и для обработки статей, вносимых в наукометрическую информационную систему авторами непосредственно в момент их добавления. В последнем случае одним из требований является достаточная производительность реализации алгоритма, позволяющая давать рекомендации пользователю непосредственно при редактировании информации о статье в интерфейсе системы. Использование хэш-функции для множества авторов статьи позволяет производить поиск возможных вариантов перевода и давать рекомендации менее чем за 0.1 сек.

ЗАКЛЮЧЕНИЕ

Разработанный нами алгоритм позволяет производить автоматический поиск переводов статей только на основе анализа графа соавторства, что обеспечивает применимость алгоритма к любым парам языков без сбора больших объемов статистической информации для построения автоматических переводов.

Результаты поиска могут использоваться самостоятельно или уточняться с использованием методов полнотекстового анализа.

Использование хэш-функции для множества соавторов позволяет значительно увеличить производительность реализации алгоритма и осуществлять поиск возможных вариантов перевода менее чем за 0.1 сек.

Результаты работы алгоритма позволяют уведомлять пользователей системы о наличии возможных переводов статей.

СПИСОК ЛИТЕРАТУРЫ

1. *Налимов В.В., Мульченко З.М.* Наукометрия. Изучение науки как информационного процесса. Москва: Наука, 1969. 340 с.

2. URL: http://www.library.spbu.ru

3. URL: http://library.bmstu.ru/Publications/

4. *Алехина Е.И.* Информационная система учета научно-исследовательской деятельности сотрудников вуза // Инновационная наука. 2018. №5-1. С. 9–12.

5. *Столяров Р.А., Чугреев В.Л.* Автоматизированная система учета результатов интеллектуальной деятельности в научной организации. URL: http://vtr.vscc.ac.ru/article/1512

6. *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.

7. *Васенин В.А., Занчурин М.А., Козицын А.С. и др.* Архитектурно-технологические аспекты разработки и сопровождения больших информационно-аналитических систем в сфере науки и образования // Программная инженерия. 2017. Т. 8. № 10. С. 448–455.

8. *Васенин В.А., Иткес А.А.* Внедрение реляционной модели логического разграничения доступа в web-приложения информационных систем, разработанных на основе библиотеки django // Программная инженерия. 2018. Т. 9. № 5. С. 195–208.

9. *Васенин В.А., Иткес А.А., Бухонов В.Ю., Галатенко А.В.* Модели логического разграничения доступа в многопользовательских системах управления наукометрическим контентом // Программная инженерия. 2016. Т. 7. № 12. С. 547–558.

10. *Коряков Д.Е.* Наукометрия. Зачем нужны разные индексы. URL: https://www.mcb.nsc.ru/sites/mcb.nsc.ru/files/fck/file/naukometriya_2.pdf

11. Автоматический переводчик «Гугл». URL: https://translate.google.ru/

12. Автоматический переводчик «Промпт». URL: http://www.translate.ru

13. Научная конференция «Обнаружение заимствований – 2017». URL: http://www.oz2017.ru

14. Плагиат в научных статьях: трудности обнаружения перевода. URL: http://ai-news.ru/2018/01/plagiat_v_nauchnyh_statyah_trudnosti_obnaruzheniya_perevoda.html

15. *Афонин С.А., Гаспарянц А.Э.* Автоматическое построение функции оценки качества в задаче разрешения неоднозначности имен авторов научных публикаций // Программная инженерия. 2015. № 10. С. 31–37.

16. Наукометрическая система «ИСТИНА». URL: https://istina.msu.ru/

ALGORITHM FOR LINKING TRANSLATED ARTICLES USING AUTHORSHIP STATISTICS

A. S. Kozitsyn[1], S. A. Afonin[2], A. A. Zenzinov[3]

Institute of Mechanics Lomonosov Moscow State University, GSP-1, Leninskie Gory, Moscow, 119991, Russian Federation

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

Abstract

During the last decades scientometric techniques have been used for research activity stimulation. Number of published articles and number of their citation counts are among the most important scientometric parameters. In an automated environment, when the publications metadata is gathered from various sources, correct linking of original papers with their translations into different languages is extremely important. In the paper we show that the known text similarity measures are inefficient in the context of article linkage problem. We propose a method for semi-automatic article linkage using statistical data on authors publication activities only. This approach may be used for linking articles without training for the language of translation. The method was evaluated on real-world collection of publications metadata of ISTINA information system.

Keywords*: bibliographic data, graph analysis, translation, article, statistics, scientometrics, citation, automated systems.*

REFERENCES

1*. Nalimov V.V., Mulchenko Z.M.* Naukometriia. Izuchenie nauki kak informatsionnogo protsessa. Moskva: Nauka, 1969. 340 s.

2. URL: http://www.library.spbu.ru

3. URL: http://library.bmstu.ru/Publications/

4. *Alekhina E.I.* Informatsionnaia sistema ucheta nauchno-issledovatelskoi deiatelnosti sotrudnikov vuza // Innovatsionnaia nauka. 2018. No 5-1. S. 9–12.

5*. Stoliarov R.A., Chugreev V.L.* Avtomatizirovannaia sistema ucheta rezultatov intellektualnoi deiatelnosti v nauchnoi organizatsii*.* URL: http://vtr.vscc.ac.ru/ article/1512

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

7. *Vasenin V.A., Zanchurin M.A., Kozitsyn A.S. etc.* Arkhitekturno-tekhnologicheskie aspekty razrabotki i soprovozhdeniia bolshikh informatsionno-analiticheskikh sistem v sfere nauki i obrazovaniia // Programmnaia inzheneriia. 2017. V. 8. No 10. S. 448–455.

8. *Vasenin V.A., Itkes A.A.* Vnedrenie reliatsionnoi modeli logicheskogo razgranicheniia dostupa v web-prilozheniia informatsionnykh sistem, razrabotannykh na osnove biblioteki django // Programmnaia inzheneriia. 2018. V. 9. No 5. S. 195–208.

9. *Vasenin V.A., Itkes A.A., Bukhonov V.Iu., Galatenko A.V.* Modeli logicheskogo razgranicheniia dostupa v mnogopolzovatelskikh sistemakh upravleniia naukometricheskim kontentom // Programmnaia inzheneriia. 2016. V. 7. No 12. S. 547–558.

10. *Koriakov D.E.* Naukometriia. Zachem nuzhny raznye indeksy. URL: https://www.mcb.nsc.ru/sites/mcb.nsc.ru/files/fck/file/naukometriya_2.pdf

11. Avtomaticheskii perevodchik «Gugl». URL: https://translate.google.ru/

12. Avtomaticheskii perevodchik «Prompt». URL: http://www.translate.ru

13. Nauchnaia Konferentsiia «Obnaruzhenie zaimstvovanii – 2017». URL: http://www.oz2017.ru

14. Plagiat v nauchnykh statiakh: trudnosti obnaruzheniia perevoda. URL: http://ai-news.ru/2018/01/plagiat_v_nauchnyh_statyah_trudnosti_obnaruzheniya_perevoda.html

15. *Afonin S.A., Gaspariants A.E.* Avtomaticheskoe postroenie funktsii otsenki kachestva v zadache razresheniia neodnoznachnosti imen avtorov nauchnykh publikatsii // Programmnaia inzheneriia. 2015. No 10. S. 31–37.

16. Naukometricheskaia sistema «ISTINA». URL: https://istina.msu.ru/

СВЕДЕНИЯ ОБ АВТОРАХ

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

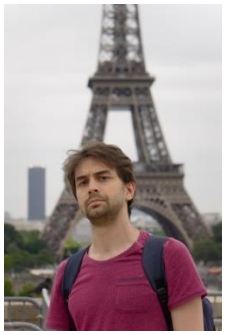
Alexander Sergeevich KOZITSYN – Leading Researcher, Ph.D., graduated from M.V. Lomonosov Moscow State University. Specialist in the field of information retrieval and database.

email: alexanderkz@mail.ru

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



ЗЕНЗИНОВ Андрей Александрович – младший научный сотрудник, окончил мехмат МГУ им. М.В. Ломоносова. Специалист в области моделирования распределённых информационных систем.

Andrey Alexandrovich ZENZINOV – Junior Researcher, graduated from M.V. Lomonosov Moscow State University. Specialist in the field of modeling of distributed information systems.

email: andrey.zenzinov@gmail.com

Материал поступил в редакцию 26 декабря 2018 года