Гуреев В.Н., Мазов Н.А. Институт нефтегазовой геологии и геофизики СО РАН Государственная публичная научно-техническая библиотека СО РАН

РЕЗУЛЬТАТЫ ИССЛЕДОВАНИЙ ПО ОБНАРУЖЕНИЮ ЗАИМСТВОВАНИЙ С ИСПОЛЬЗОВАНИЕМ АНАЛИЗА ЦИТИРОВАНИЙ

Определение

Переводной плагиат – перевод текста с одного языка на другой без отсылки к первоисточнику и его последующее опубликование. Является наиболее сложным для выявления типом плагиата.

Отличительные особенности

- распространен в научной сфере, где рецензирование строже и простой плагиат быстро выявляется;
- распространен в неанглоязычных странах, где ученые:
 - ✓ переводят англоязычные тексты и публикуют их на своих языках;
 - ✓ переводят собственные работы с родного языка на английский (самоплагиат);
- исследователи отмечают, что росту популярности способствуют успехи машинного перевода.

Начало исследований

Интерес к выявлению переводного плагиата возник в начале 2010-х гг.



China's Publication Bazaar

systems medy: "Pattink Mr. Piges", with the mittors line. "Dame for co-flor series and co-corrupteding author." A resonage body metabolical with physical for commend a mixed to planter with all his second improducess, had once surface customs. This discussion would find some rolls. Set her surpose in Christia hermanisad have all the set of the se

the their profit shot, and they get he requisited in the con-traction for the contract of the

procedure of the control of the cont The basis little can be left to us. Our nevent can halv two. Asset \$120

September, resplity when the agent said it would. The total and seperhald addresses one expression from the proof solutions: that the last of polices was transfermed. On the periodical paper,

them Nimery, an incrementary by the Range and Joseph's Such sectors of Call Blad per Based that a total of Base pathete had been pilled, and two dispipal. The enthylene agent had section was detail disting out have consequence with fact. To those, the pages the paper, throughing a project of strategy the melting they constance. And prolesymes was mixed of covers, but had not yet been assumed

ments and appropriate modern to make extillustrate. The company and Marshaul Basis before after the other propriate articles, which

The solid security office to be from an assumpty to he basely substituted in the control of the terminal size being your against 4 pills. The mater against a riber to the transcription of the program is reasonable program. It is control of the program in the control of the program is the control of the program in the program in the program is the program in the program in the program is the program in the program in the program is the program in the program in the program is the program in the program in the program is the program in the program in the program in the program is the program in the program in the program in the program in the program is the program in the program in the program in the program in the program is the program in the program in the program in the program in the program is the program in the program i no first authors show the former of he separate that not pay for to \$35,500, At the high and, fors record the annual subsets of none unbeddigt. Deprises with orders and with the journal's offense. Deback sociated politicism. See SET paper—pointed this continued that a level author to a dated of 11 Janua, agreementary political of a journal with a high depart factor on to called the visibilities or pointer record. He deback of the factor between printing permanents and are according to the deback of the printing permanents and are according to the deback of the printing permanents and are according to the printing permanents.

Comparative Evaluation of Text- and Citation-based Plagiarism Detection Approaches using GuttenPlag

OvDU, Bernary & UC Burkeley

Norman Meuschie CVCE: Germany & UC Berneley

Joeran Beel OvSU Germany & UC Berkelov

Tomas appearance for projection describes state. All one based on scott or less against state analysis puriods such as entiry copoling, Engagementing of cryb comparison. So this pages a new gramech called Citation-Head Physicists Distuition is evaluated operant acted Chains-band Pagama Basiline is column tongs a death from [1] is the fields a relative promotioning proper called Goldenfing [1] identified infrastrict enterest of plagation foreign cartill minute suppose. The see appoint is able to detailly called and disputated documents haved on the statebook mode on the next 8 to dense that plastimateurs. recolors in Mediting strong peoplewing, remission and other 200 plagration. Detection many can be improved by continuing station bound with soil based plagration (Secretors)

Categories and Subject Descriptors

8.1.1 (Cheering) PATOESLATERS STORAGE AND RETREVAL Spreadure leave and Reserval General Terms

Paperer Dennie Spore, Chica-back Reporter Dencies

I. INTRODUCTION

exhibited or country with our passing flags off at our town to the R. S. S. positionarily represent problem among coffage students. but also provides among contributed returnings. [17]

Booods, a plaqueism cycle tendring the German research of delines. No. Genturines, getted relationard profess mention. For chance, a line postered should proposition become white No. Hallanders, No. Americal States (3). More For proposite politicals republished for accountance, or "diversor", orderment sections for Controlling proper [1] to development the torontymes and Sommer the true around of physicism promed in So. work.

As all April 10° 2011, the jone offers remarked that 211 out of 191 main text pages to the theory conducted plaquetized thapmone. In radio shout 64°, of the yeal flows now bines had as plaquement. The fittings have been

26 50 50 20 20 20 30 50 40 40 Figure 5 Pages containing Plagfacture (sancture (15))

source, not free proger with phagastree friend analysis; asserted and As a count of the analysis, NA. Training no large cleaned his facility to be flowing. By enterprised his discusse and eventually support drives then his political position.

Takashang's dissert on other to discontinuo on possibili for dissistivity, cliently, dispeted diagnosis, respectivity to some of dispersion has it admin translated plaganeses. New princy or currency of different tenso of plaganeses and the centrally used distriction conflicts, fist, concept of Catrion-based Physicianus (Antonios (1997s) is dissisted. Alternatio, fist conflictivity; the evaluation is presented, its fire fluid vertices, the advantages in

Effectivation of physician behavior in gracine every a resofter of commonly from earliest for illustrate test sugge, which can be selly be summarized at fiftees. Caped from challen plays to specifies the six of subangerous parts of the society of a lost reference from another solitor. Physician Juggierous includes grantes stended to mak listedly contol segments. Under practice strends to come model; compay augments. I reac-incapitations of children for controllar levelings of Disciplina, in the conductory and right of the plaquistic estimat piring that could be estate to contain the neighbor among 10°, Franchison, and the controllar to contain the neighbor among 10°, Franchison, plaquism in the casual or introductal concentration of contains these are transpired on among the controllar as some the straight. Over pliquinto coccustamo de suaje el intrader foriga coscoli without agregoram musus administrativamos. As Excepti o de appropriation el research agressimo, enclusis, papersonani scope, a generalisto o discrete. Balliguand contro de 191

2.2 Phylarius Detection Approaches

Register Develop (PD) is a bootset for company has Program on Processing by Sections of Plant is a composed theory procession is approximately the Sections of of plant point in evidence Exhaust PM systems OFGE can be comparated one charged and statement. Electron PMI compare a management advantage to entire of generate works. Sections 1932 systems only specials the plant of generate works. Sections 1932 systems only specials the plant of generate works. Sections 1932 systems of general designation terms of the configuration 1932, a principle form on the plant of the plant of the configuration 1932. chy without perferring companions to entered and, Misle counter PDS are to fast brands starting and m, artissos PDS try to succepture thanges to string orde (1.7).

Citation Analysis as a Basis for the Development of an Additional Module in Antiplagiarism Systems

V. N. Gureev* and N. A. Masov*

"Figure State Research Gener of Verlagy and Born hoology, Novemberk white", Kal'isone, Russia. a-mail present Charge as as *Traffmak Eustines of Persolvan Geology and Geophysics, Schoolar Bounds, Eustine Analony of Sciences, Nasoublink, Rassia ar-most: Messer NA Higgs Johnson Received August 26, 2013

Notified ... The one of lists of references in different arous is proving exponentially. Sowing as bain for defining the intrinsmidige between earthure, publications, and journals, informace good an impures to the develop-ment of a set of indicators and indicate and the all a large impact on the development of bibliometries and secretarizations, in this dural paper the use of informace as an adultional models of an intelligengation system. to suggested. It is shown how elected betweening from executing works out to intentified by comparing bibliographics to view of their unique structure, which varies from one publication to another.

Krywest: plagamen, citation analysis, citation models, scientific introlation, Williagraphic dandous DOI: 10.3101/5014768213040151

INTRODUCTION

The growing number of scientific publications and an increase in the number of researchers worldwide is accompanied by a parallel increase in cases of plagarium (bornneings of somoune's test without oriening to the original source). Special withware for plagiartor dotaction is developed in order to commensal multiractice by uncrepulous authors (e.g., 11)).

To date, both free software of literack functionality and expensive powerful systems for detecting physic mm in disentations, which are prochased by large universities and penters, have been speaked. All of three orders are built on the basis of automated text-indexing algorithms. Such uniters allow one to relatively efficiently identify plagrarized fragments of work and to show the share of the original text. It is not the aint of this paper to review these particular systems Olsey were presented in detail in [2]). Rather, the observing is to highlight one of the proferm that has sol yet been modified the operability of such systems is firefied to working with only one language and an explicitly copied test. Given the fact that most of the exestific literature is published in English, Englishspeaking authors from non-English speaking coentries sorontimus simply translate one or accord publications for their subsequent compilation in order to publish this work as an original soudy in their native language. One should note that the first type of plugiorism, which does not issolve linguistic translation, is mainly widespread among students. The second type of plugiarium, which is a more complex one, largely

affects the research sector. Plugarism of sites is less common, but is rather difficult to identify.

Due to the generally higher level of scientific micarch, impalated publications more quickly go through poor review in national journals and are subsequently more frequently cited. The translated works are also used for dissertations and reports prepared as port of grants and publicly funded federal research programs. It should be record that researchers often refuse to computer translation as plagiarion, preferring to call it "abundant citation" [7], especially in those cases where papers provide a general reference to the original source at the fregioning. The publication of such works is justified by the lack of ingle-quality Res-stan-language publications in a sumber of scientific disciplines. It is, however, clear that such works do not create any new knowledge and should be published as translations, which are certainly highly needed.

At present, it is only possible to identify translation-hazed plagtomen or plagtonism of ideas by engaging highly qualified professionals familiar with a broad range of scientific publications in their area of research. Obviously, such a type of analysis can only be applied on a cone by case basis given its high favancial and labor costs. Uncompalisas authors, who fast relatively safe and opt for less time-comparing translation of foreign works inwead of carrying out independent publicly funded research, take advantage of this situation. At a result, the number of volumble domestic research developments shrinks; studies that have already been conducted by foreign counterparts and

Современные способы выявления

- Научная экспертиза требует привлечения большого числа ученых, что затратно по времени и оплате; неприменима для массовой обработки подозрительных текстов.
- Выявление плагиата с применением анализа цитирования, при котором два текста разных авторов со схожими списками литературы дают основание предполагать наличие плагиата.
- Автоматизированные системы прямого сличения текстов на разных языках только начинают появляться.

Обоснование метода

При переводе текста на родной язык авторы сохраняют всю или значительную долю ссылок, а нередко и их последовательность.

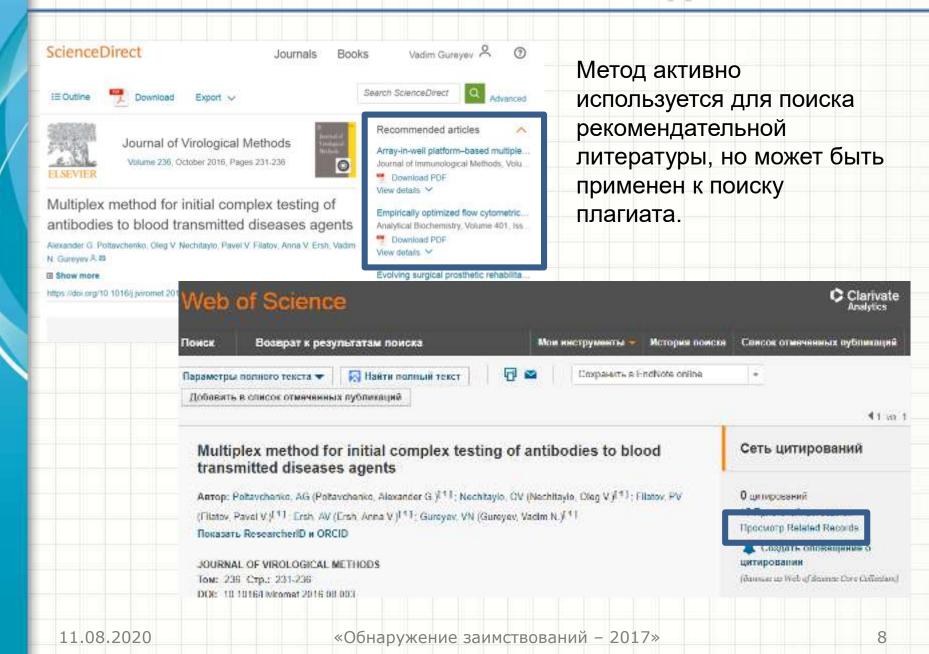
Обнаружив в двух текстах идентичные списки процитированной литературы, можно предположить, что более поздняя публикация является плагиатом.

Обоснование метода

Преимущества использования цитирований перед прямым сличением текстов в том, что цитирования не зависят от языков и в то же время являются очень емкими поисковыми образами (Ю. Гарфилд), доказывающими семантическую близость двух текстов.

Метод основан на понятии библиографического сочетания (bibliographic coupling) – работы М. Кесслера начала 1960-х гг.: две работы имеют осмысленное отношение друг к другу и тематически связаны, если у них есть одна и более общих ссылок в пристатейных списках литературы.

Обоснование метода



Основной вопрос

Основная проблема заключается не столько в самом сличении цитирований двух текстов, сколько в поиске возможного оригинала, имеющего схожий список литературы с анализируемой публикацией.

Подходы к поиску оригинального текста

Оригинал для переведенного текста может быть найден:

- в полнотекстовой базе данных методом семантического анализа (система тестируется в «Антиплагиате»)
- в полнотекстовой базе данных методом анализа цитирования (коллектив под руководством Б. Гиппа)
- в библиографической базе данных методом анализа цитирования (наша разработка)





- дороговизна подписки электронных ресурсов разных издательств;
- зависимость от полных текстов;
 - невозможность охватить все источники.

доступность библиографических БД во всех российских организациях;

- дешевизна фактологической базы;
- очень широкий охват источников.

Использование полнотекстовых БД

Преимущества:

- анализ как набора ссылок, так и их последовательности;
- возможность проведения семантического анализа.

Недостатки:

- отсутствие всеохватной базы данных с полными текстами;
- дороговизна подписки из-за платного доступа к большинству полнотекстовых баз данных.

Использование библиографических БД

Преимущества:

- использование библиографических БД дает возможность поиска оригинала в очень широком диапазоне научных текстов;
- возможна автоматизированная генерация запроса в БД на основе цитируемых в подозрительном тексте источников;
- имеющийся инструментарий БД дает широкие возможности поиска по пристатейным спискам литературы.

Недостатки:

- невозможен анализ последовательности ссылок в случае приведения ссылок в оригинале по алфавиту, т.к. нужен доступ к полному тексту;
- невозможность семантического анализа текстов.

Ограничения метода

Текущие ограничения анализа цитирования:

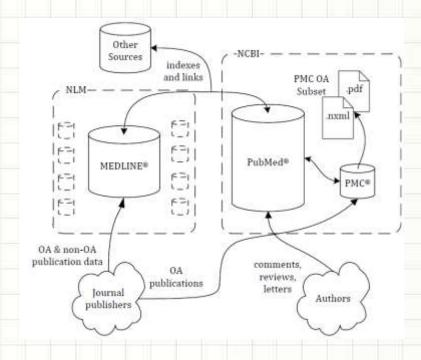
- малое число ссылок;
- цитирование литературы, плохо индексируемой в базах данных (характерно для общественных и гуманитарных наук).

Ограничения, которые уже нашли разрешение:

- перемешивание ссылок;
- масштабирование ссылок (большее или меньшее число использований в сравнении с оригиналом);
- пропуск ссылок;
- добавление собственных ссылок.

Апробация метода в полнотекстовых БД

Масштабная проверка метода с применением анализа цитирования проводилась коллективом Б. Гиппа с использованием полнотекстовой базы данных PubMed Central Open Access Subset, где была проведена полная проверка каждого текста со всей совокупностью других текстов (более 185 тыс., из которых выявлено несколько десятков случаев плагиата).



Апробация метода в полнотекстовых БД

Прототип программы (www.citeplag.org) на основе анализа цитирования группы Б. Гиппа позволяет сравнивать лишь два имеющихся текста.

Возможность самого поиска оригинала в настоящий момент не реализована.

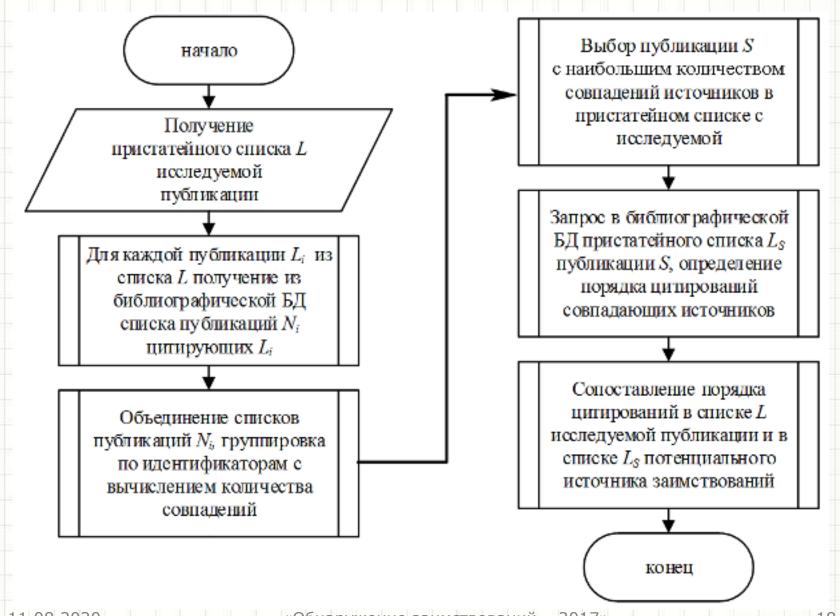
Апробация метода в библиографических БД

В режиме создания и адресации запросов в мультидисциплинарные БД Scopus и Web of Science нашим коллективом было подтверждено несколько случаев плагиата в области информатики и биологических наук.

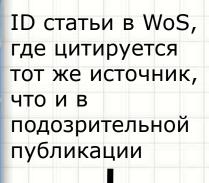
Алгоритм поиска оригинала

- 1. для каждого источника в списке литературы подозрительной публикации создавался запрос в библиографическую БД;
- 2. выявлялась группа публикаций, цитирующих тот же источник;
- 3. выгружался список идентификаторов публикаций, цитирующих те же источники, что были в списке литературы подозрительной публикации;
- 4. напротив идентификаторов проставлялись порядковые номера источников из списка литературы подозрительной публикации.

Схема алгоритма



Пример



Номер цитируемой статьи в списке литературы подозрительной публикации

000078603800016***58 000078603800055***18 000078603800077***12***18***19 000078614200002***43***44 000078640400003***43

Совпадение двух или трех источников не вызывает подозрений

000278518900002***38***45***46***47***48***49***50***51***5
2***53***54***55***56***57***58***59***60***61***62***63

ID искомого оригинала, где последовательно цитируются сразу 18 источников из списка литературы подозрительной статьи

Пример

A SECRETARY OF THE PARTY OF THE

combinate the course of the course for the course of

CHARLE ME I ARREST STORY AND ARREST STORY

Русскоязычная статья

Для запуска адаптивного иммунного ответа против ВГ необходимы профессиональные антиген-презентирующие клетки (АРС), такие как дендритные клетки (DC). Уменьшение численности DC в легких в процессе инфекции приводило к снижению уровня ИФН 1-го типа и СВ8-клеток [53]. Альвеолярные макрофаги и ДК одними из первых активируются инфекцией ВГ типа А [54]. Во время заражения альвеолярные макрофаги производят большое количество воспалительных цитокинов, главным образом интерлейкина-6 (ИЛ-6) и фактор некроза опухолей-и (ФНО-и). Кроме того, инфицированные эпителиальные клетки легких выделяют в больших дозах моноцитарный хемоаттрактантный белок 1 (МСР-1), мобилизующий моноциты, которые могут видоизменяться в DC и свободные макрофаги [55]. Во время активации DC созревают и мигрируют в дренирующие лимфатические уалы легких, где они представляют пептидный антиген ВГ в комплексе с молекулами МНС класса I предшественникам CD8°-Т-лимфоцитов. Это вызывает пролиферацию и дифференцировку Т-клеток в эффекторные клетки (цитотоксические CD8"-Т-лимфоциты).

В процессе зволюции ВГ приобрел механизмы, которые способны подавлять иммунитет хозяина. В частности показано, что белок NS1 ограничивает продукцию ИФН 1-го типа, а также других цитожинов и хемокинов во время заражения вирусом гриппа А [56]. Более того, ВГ способен предотвращать соэревание DC. Зараженные DC имеют сниженную экспрессию молекул МНС класса 1 и таких костимулирующих рецепторов, как СD80 и CD86, что прелятствует представлению эндогенно экспрессируемых вирусных антигенов CD8-Т-клаткам [57].

Список литературы русскоязычной статьи

induction of antivital responses in primary human dentifits and respiratory applicable cells, J Virol. 2009;83:6649-62.

 Young LJ, Wilson NS, Schnorer P, Mourt A, Lunde RJ, La Guss ML, et al. Denditic cell preschivation impain: WHC class is presentation of vaccines and analogenous viral antigens. Proc Natl Acad Sci USA. 2007;104:17753-8.

K MARKET THE THE BANK A NEW YORK TO BE

A STREET WAS AS ADDRESS OF THE PARTY OF THE

Married Street, Square Str., Str. or Str. or

STATE OF STREET STATE OF STREET STREET

 McGR J. Heusel JW, Legge KL. Innuts intrium control and regulation of influenza sinus offictions. J Leukoc Biol. 2009;88:1003–12.

man on a single or man from the first the

- Michalla JM, Chan MC, Chan WY, Wong HK, Cheung DY, Keong DL, et al. Tropiers of areas officercis A (HSW1) in the upper and lower respectory tract. Nat Med. 2007;12:147. A
- Herold S, von Wadfee W, Steinmunike M, Floodfac S, Kuzel WA, Mack M, et al. Advance optimized lists descriptions/per transportation registroscopic influences after influences. Impact of chemokines and adhesion malecules. J Internact. 2006;177:1987-24.
- Haye K, Burmakina S, Morae T, Garcia-Sastre A, Fersandez-Sesma A. The MS1 protein of a human influency virus inhibits type I interferon production and the

The quality of the initial inhale response against IAVs has profound consequences for the following adaptive response. Triggering the adaptive immune response against IAVs requires professional APC-like DCs. Depletion of lung resident DCs was shown to result in reduced requirement of pDCs (reviewed in NoSiLef at 2008s), which are the main producers of type LIFN, and CD6+ cells upon IAV intection.

Residential alveolar macrephages and dendrific cells are among the first cells activated by I/V Infection. Depending on the virus strain a significant proportion of infected cells are shedlar macrophages (Michaels et al., 2007). Upon infection with I/Vs alveolar macrophages produce is guideant amounts of inflammatory cytotines like IL-6 and TNF or and become highly phagocytic. Additionally, infected lung epithetial cells secrets high levels of monocyte dremnettractant protein 1 (HCP-1), recruiting monocytes, which can differentiate into monocyte derived DCs and inflammatory medicohages (Herod et al., 2006).

Upon activation, DCs undergo maturation and migrate to the draining lymph nodes of the lung where they present peptide antigens to IAV-specific naive Tipets. This results in profiferation and differentiation of these Tipets into effection cells. As shown for other cell types, NS1 limits the production of type ITEM as well as that of cytokines and chemokines upon IAV irrection (<u>Ferrander-Germaler at 2006</u>. <u>Have et al. 2008</u>). Horeever, IAVs are capable of interfering with the maturation of DCs, infected DCs show reduced upregulation of IVEM class II and costimulatory receptors into CD80 and CD85, essential for 1 cell printing. Young and colleagues demonstrated that infection of mature DCs by influence with the prepentation of an endogenous viral antigen and can thus impact the people immune response. Young et al., 2007).

Англоязычный оригинал

Пример

Наиболее интересный и сложный случай – отчет в один из грантовых фондов

- 202 ссылки в списке литературы;
- 194 источника нашлись в Scopus;
- на эти 194 источника также ссылались 29971 раз из 17228 публикаций.

Распределение ссылок:

- по 1 совпадению 12711 публикаций;
- 2-5 совпадений 4440 публикаций;
- 6-10 совпадений 442 публикации;
- 11–20 совпадений 108 публикаций;
- 21-30 совпадений 18 публикаций;
- 31-40 совпадений 6 публикаций.

Анализ результатов:

отчет составлен из переводов трех статей с числом совпадений ссылок 41, 65 и 82.

Текущее состояние проекта

- Проводятся работы по автоматизации генерации строк запросов в БД и обработки выгруженных из БД в текстовый формат списков литературы.
- Высокая степень автоматизации может быть достигнута взаимодействием с веб-интерфейсами БД с имитацией деятельности пользователя (например, программным интерфейсом Scopus Search API).
- Публичный доступ к сервису поиска потенциальных источников заимствований по приведенному алгоритму с использованием АРІ будет нарушать политику использования программных интерфейсов правообладателя и потребует специального согласования, которое может быть достигнуто крупными системами выявления плагиата.

Финансирование

В 2016 г. наша разработка модели выявления плагиата на основе анализа цитирований получила поддержку РФФИ (проект 16-07-00652). Проект продолжается.

Спасибо за внимание!

Гуреев Вадим Николаевич, с.н.с. ИАЦ ИНГГ СО РАН; с.н.с. ГПНТБ СО РАН

Мазов Николай Алексеевич, зав. ИАЦ ИНГГ СО РАН; с.н.с. ГПНТБ СО РАН

E-mail: <u>GureyevVN@ipgg.sbras.ru</u> <u>MazovNA@ipgg.sbras.ru</u>