УДК 001.893:347.77/.78:001.811:004.65

РЕЗУЛЬТАТЫ ИССЛЕДОВАНИЙ ПО ОБНАРУЖЕНИЮ ЗАИМСТВОВАНИЙ С ИСПОЛЬЗОВАНИЕМ АНАЛИЗА ЦИТИРОВАНИЙ

В.Н. Гуреев[1,2], Н.А. Мазов[1,2]

[1]*Институт нефтегазовой геологии и геофизики им. А.А. Трофимука СО РАН, 630090, Новосибирск, пр. Акад. Коптюга, 3;*
[2]*Государственная публичная научно-техническая библиотека СО РАН, 630200, Новосибирск, ул. Восход, 15*

GureyevVN@ipgg.sbras.ru, MazovNA@ipgg.sbras.ru

Аннотация

Переводной плагиат как одна из наиболее распространенных в научном информационном пространстве разновидностей плагиата представляет собой трудноразрешимую проблему, поскольку практически не поддается автоматизированному выявлению. Между тем за последние пять лет в этом направлении наблюдается прогресс. Авторами настоящей работы, а также группой зарубежных исследователей из нескольких университетов независимо друг от друга был предложен подход к выявлению плагиата на основе анализа цитирований, при котором для анализируемой подозрительной публикации находится возможный первоисточник с идентичным или схожим списком цитируемой литературы, что в итоге позволяет сличать текст на разных языках. Разработанная методика обнаружения неправомерных заимствований в научных текстах успешно прошла тестовые исследования. В статье приведены результаты четырехлетних исследований.

Ключевые слова: обнаружение заимствований, переводной плагиат, выявление плагиата, анализ цитирования, база данных цитирований

ВВЕДЕНИЕ

Метод анализа цитирований в научных публикациях имеет множество различных практических приложений. Одной из последних разработок, параллельно проводимой авторами настоящей работы [1–3] и коллективом зарубеж-

ных исследователей под руководством Б. Гиппа [4–6], является использование анализа цитирования в применении к выявлению плагиата. Данный метод позволяет оценить количество общих цитирований в сравниваемых публикациях, порядок их появления, близость расположения друг к другу в тексте и вероятность их совместного появления. Особенность и преимущество данного подхода в сравнении с другими методами выявления плагиата заключается в независимости анализа от лексических совпадений в сопоставляемых текстах. При анализе цитирования обрабатываются лишь списки цитируемой литературы и их последовательности, что позволяет абстрагироваться от самих текстов публикаций и в итоге обойти пока не разрешимую проблему сличения текстов на разных языках.

1. АНАЛИЗ ЦИТИРОВАНИЯ ДЛЯ ВЫЯВЛЕНИЯ СЛУЧАЕВ ПЕРЕВОДНОГО ПЛАГИАТА

Основными преимуществами использования анализа цитирования для выявления случаев переводного плагиата являются следующие:

а) списки литературы – обязательный атрибут научных текстов, что дает возможность их сравнивать;

б) списки литературы общедоступны, в большинстве случаев бесплатны, а их анализ не требует доступа к полным текстам;

в) сравнительный анализ списков литературы можно автоматизировать, что делает возможным массовую проверку текстов.

Возможными проблемами на пути выявления первоисточников могут быть, во-первых, малое число ссылок в тексте, а во-вторых, усилия плагиатора. В частности, ссылки могут быть перемешаны, а их конечное число может быть масштабировано в большую или меньшую стороны.

Вслед за завершенным теоретическим обоснованием возможности использования данного подхода в качестве дополнительного модуля в системах выявления плагиата [2, 7] встала задача автоматизации процессов поиска возможных оригиналов при анализе подозрительных публикаций. Для успешного решения задачи выявления плагиата, основанного на сличении моделей цитирования в двух текстах, мы использовали достижения библиометрии, а именно, исследования, касающиеся библиографического сочетания, или библиографиче-

ской связи (bibliographic coupling), предложенного М. М. Кесслером [8, 9]. При библиографическом сочетании за единицу связывания между двумя статьями принята общая ссылка из двух публикаций. Такие две статьи считаются библиографически связанными. Сила их библиографического сочетания, таким образом, – это количество общих для них ссылок.

Метод кластеризации результатов библиографического запроса, основанный на библиометрическом методе библиографического сочетания, предполагает, что две работы имеют осмысленное отношение друг к другу и тематически связаны, если у них есть одна и более общих ссылок в пристатейных списках литературы [8, 9]. Основными преимуществами данного метода являются его независимость от лексики и языка публикаций, а также возможность автоматизации. Поэтому он напрямую может применяться также к проблеме выявления переводного плагиата. В данном случае более поздняя работа, имеющая сильно схожий список литературы с более ранней публикацией, независимо от языка документов, становится объектом анализа на возможные заимствования.

2. КОНЦЕПЦИЯ ИСПОЛЬЗОВАНИЯ МУЛЬТИДИСЦИПЛИНАРНЫХ БИБЛИОГРАФИЧЕСКИХ БАЗ ДАННЫХ ДЛЯ АНАЛИЗА ЦИТИРОВАНИЯ

В настоящее время во всех существующих программных решениях по выявлению заимствований в публикациях необходим доступ к полным текстам, поскольку сличаются именно тексты публикаций (см., например, [10, 11]). Для наиболее точного выявления переводного плагиата с помощью анализа цитирования также желательны полные тексты, поскольку при построении списка литературы в алфавитном порядке невозможно восстановить последовательность цитирований.

Основным недостатком в подходе, предложенном зарубежными исследователями, является зависимость от наличия полных текстов для анализа списков цитирований. Нами был предложена концепция использования мультидисциплинарных библиографических баз данных для анализа цитирования для выявления плагиата. В качестве библиографической базы данных может использоваться любая система с поддержкой возможности просмотра списков цитировавших публикацию работ.

Принципиально алгоритм состоит из следующих шагов:

1. Для каждого цитируемого источника из списка литературы подозрительной публикации формируется запрос в библиографическую базу данных с целью извлечь список публикаций, также цитировавших этот источник.
2. Полученные для каждого цитируемого источника списки объединяются с подсчетом количества совпадений. Таким образом, мы получаем список публикаций, цитировавших те же источники, что и подозрительная работа, с указанием количества совпадающих источников. Ранжированный по убыванию количества совпадающих источников список публикаций является предметом дальнейшего анализа на некорректные заимствования. В зависимости от обстоятельств исследования этот список может быть сокращен отсечением по абсолютной (например, количество совпадающих источников более 8) или относительной (например, количество совпадающих источников более 40 % от всего списка литературы исследуемой статьи) границе.
3. В случае, если в исследуемой работе список литературы организован в порядке цитирования, формируется дополнительный запрос к библиографической базе данных для получения списка литературы потенциального источника заимствований. Если этот список также организован в порядке цитирования, проводится анализ совпадения порядка цитирований.

Конкретными результатами исследований, проведенных нашей группой, стало выявление плагиата в различных типах научных публикаций. Были проверены некоторые монографии, научные статьи, отчеты о НИР и диссертационные работы. Метод тестировался в двух мультидисциплинарных базах данных – Web of Science Core Collection и Scopus. Обе базы данных на основе интерфейса API позволяют автоматизировать запросы по спискам литературы и получить список публикаций, ссылающихся на те же источники, на которые сделал ссылки автор подозрительной публикации. Нами были выявлены случаи плагиата в научных статьях, одной монографии, главы которой были переведены из разных статей, а также в одном отчете о НИР, который был в нашем распоряжении. Результаты по отчету о НИР представляют особенный интерес, поскольку работа проводи-

лось на очень обширном ссылочном аппарате, что во многом позволило уточнить методику. Отчет содержал 202 ссылки, 194 из которых были проиндексированы в Scopus. На эти 194 работы была сделана 29971 ссылка из 17228 публикаций. Распределение ссылок было следующим: по 1 совпадению было в 12711 публикациях, от 2 до 5 совпадений – 4440 публикациях, от 6 до 10 совпадений – в 442 публикациях, от 11 до 20 совпадений – в 108 публикациях, от 21 до 30 совпадений – в 18 публикациях, от 31 до 40 совпадений – в 6 публикациях. В трех случаях было по 41, 65 и 82 совпадениям. Анализ данных показал, что отчет был составлен из переводов трех статей, где было по 82, 65 и 36 совпадений. Дальнейшее исследование публикаций с подозрительно высоким числом совпадений, в частности, работы с 41 общей ссылкой, дало дополнительным результатом выявление еще одного случая самоплагиата в зарубежной статье одного автора.

ЗАКЛЮЧЕНИЕ

Автоматизация разработанной нами модели выявления плагиата впоследствии может эффективно применяться для определения неправомерных случаев текстовых заимствований. Алгоритмы, заложенные в модели, могут быть применимы непосредственно в компьютерных программах для автоматизации поиска оригинальных текстов и визуализации полученных результатов, что входит в наши дальнейшие планы работ в данном направлении. Разработка и промышленный запуск подобной системы позволили бы, на наш взгляд, значительно снизить объемы заимствований.

Благодарности

Исследование выполнено при финансовой поддержке РФФИ в рамках научного проекта № 16-07-00652.

СПИСОК ЛИТЕРАТУРЫ

1. *Mazov N.A., Gureev V.N., Kosyakov D.V.* On the development of a plagiarism detection model based on citation analysis using a bibliographic database // Scientific and Technical Information Processing. 2016. V. 43, No 4. P. 236–240.

2. *Gureev V.N., Mazov N.A.* Citation analysis as a basis for the development of an additional module in antiplagiarism systems // Scientific and Technical Information Processing. 2013. V. 40, No 4. P. 264–267.

3. *Мазов Н.А., Гуреев В.Н.* К вопросу о разработке моделей выявления плагиата на основе цитирования с использованием наукометрических баз данных // Библиотеки и информационные ресурсы в современном мире науки, культуры, образования и бизнеса: Труды 23-й Международной конференции «Крым-2016» (4–12 июня 2016 г., Судак). М.: Изд-во ГПНТБ России, 2016. С. 1–4.

4. *Gipp B., Meuschke N., Breitinger C., Lipinski M., Nürnberger A.* Demonstration of citation pattern analysis for plagiarism detection // 36-th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2013 (July 28 – August 01, 2013, Dublin, Ireland). New York: ACM, 2013.P. 1119–1120.

5. *Meuschke N., Gipp B.* Reducing computational effort for plagiarism detection by using citation characteristics to limit retrieval space // Proceedings of the ACM/IEEE Joint Conference on Digital Libraries, 2014. P. 197–200. Doi: 10.1109/JCDL.2014.6970168.

6. *Gipp B., Meuschke N., Breitinger C., Pitman J., Nürnberger A.* Web-based demonstration of semantic similarity detection using citation pattern visualization for a cross language plagiarism case // ICEIS 2014. Proceedings of the 16th International Conference on Enterprise Information Systems. 2014, V. 2, P. 677–683.

7. *Gipp B., Meuschke N.* Citation Pattern Matching Algorithms for Citation-based Plagiarism Detection: Greedy Citation Tiling, Citation Chunking and Longest Common Citation Sequence // Proceedings of the 11-th ACM symposium on Document engineering (DocEng '11) (19–22 September, 2011, Mountain View, USA). New York: ACM, 2011. P. 1–10.

8. *Kessler M.M.* An Experimental Study of Bibliographic Coupling Between Technical Papers // IEEE Transactions on Information Theory. 1963. V. 9, No 1. P. 49–51.

9. *Kessler M.M.* Comparison of the results of bibliographic coupling and analytic subject indexing // American Documentation. 1965. V. 16, No 3. P. 223–233.

10. *Осипов Г.С., Смирнов И.В., Тихомиров И.А., Соченков И.В., Зубарев Д.В., Исаков В.А.* Технологии семантического поиска заимствований в научных текстах // Труды 23-й Международной конференции «Библиотеки и информационные ресурсы в современном мире науки, культуры, образования и бизнеса» (4–12 июня 2016 г., г. Судак). М. : ГПНТБ России, 2016. С. 1–3.

11. *Sochenkov I., Zubarev D., Tikhomirov I., Smirnov I., Shelmanov A., Suvorov R., Osipov G.* Exactus Like: Plagiarism Detection in Scientific Texts // Advances in Information Retrieval: 38th European Conference on IR Research, ECIR 2016 (March 20–23, 2016, Padua, Italy). Cham: Springer International Publishing, 2016. P. 837–840.

STUDY RESULTS FOR THE DETECTION OF MATCHING CONTENT USING CITATION ANALYSIS

V.N. Gureyev[1,2], N.A. Mazov[1,2]

[1] Trofimuk Institute of Petroleum Geology and Geophysics, SB RAS, Koptug ave. 3, Novosibirsk, Russia, 630090;

[2] State Public Scientific Technological Library, SB RAS, Voskhod ave., 15, Novosibirsk, Russia, 630200

GureyevVN@ipgg.sbras.ru; MazovNA@ipgg.sbras.ru

Abstract

Translated plagiarism has widely spread in a scientific world and posed a serious problem due to the challenges in its automatic detection. However, in the last five years some progress has been observed in this area. The authors of this paper, as well as foreign research team from several universities independently of each other proposed an approach to detect plagiarism based on citation analysis with search of initial source for analyzed suspected paper with the same or similar references. Developed methods of detection of illegal use of borrowed text successfully passed several tests. The report shows the results that we have obtained in the last four years.

Keywords: detection of matching content, translated plagiarism, plagiarism detection, citation analysis, bibliographic database

REFERENCES

1. *Mazov N.A., Gureev V.N., Kosyakov D.V.* On the development of a plagiarism detection model based on citation analysis using a bibliographic database // Scientific and Technical Information Processing. 2016. V. 43, No 4. P. 236–240.

2. *Gureev V.N., Mazov N.A.* Citation analysis as a basis for the development of an additional module in antiplagiarism systems // Scientific and Technical Information Processing. 2013. V. 40, No 4. P. 264–267.

3. *Mazov N.A., Gureev V.N.* K voprosu o razrabotke modelej vyyavleniya plagiata na osnove tsitirovaniya s ispol'zovaniem naukometricheskih baz dannyh // Biblioteki i informatsionnye resursy v sovremennom mire nauki, kul'tury, obrazovaniya i biznesa: Trudy 23-j Mezhdunarodnoj konferentsii «Krym-2016» (4–12 iyunya 2016, Sudak). M.: GPNTB Rossii, 2016. 4 p.

4. *Gipp B., Meuschke N., Breitinger C., Lipinski M., Nürnberger A.* Demonstration of citation pattern analysis for plagiarism detection // 36-th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2013 (July 28 – August 01, 2013, Dublin, Ireland). New York: ACM, 2013.P. 1119–1120.

5. *Meuschke N., Gipp B.* Reducing computational effort for plagiarism detection by using citation characteristics to limit retrieval space // Proceedings of the ACM/IEEE Joint Conference on Digital Libraries, 2014. P. 197–200. Doi: 10.1109/JCDL.2014.6970168.

6. *Gipp B., Meuschke N., Breitinger C., Pitman J., Nürnberger A.* Web-based demonstration of semantic similarity detection using citation pattern visualization for a cross language plagiarism case // ICEIS 2014. Proceedings of the 16th International Conference on Enterprise Information Systems. 2014, V. 2, P. 677–683.

7. *Gipp B., Meuschke N.* Citation Pattern Matching Algorithms for Citation-based Plagiarism Detection: Greedy Citation Tiling, Citation Chunking and Longest Common Citation Sequence // Proceedings of the 11-th ACM symposium on Document engineering (DocEng '11) (19–22 September, 2011, Mountain View, USA). New York: ACM, 2011. P. 1–10.

8. *Kessler M.M.* An Experimental Study of Bibliographic Coupling Between Technical Papers // IEEE Transactions on Information Theory. 1963. V. 9, No 1. P. 49–51.

9. *Kessler M.M.* Comparison of the results of bibliographic coupling and analytic subject indexing // American Documentation. 1965. V. 16, No 3. P. 223–233.

10. *Osipov G.S., Smirnov I.V., Tikhomirov I.A., Sochenkov I.V., Zubarev D.V., Isakov V.A.* Tekhnologii semanticheskogo poiska zaimstvovanij v nauchnyh tekstakh // Trudy 23-j Mezhdunarodnoj konferentsii «Biblioteki i informatsionnye resursy v sovremennom mire nauki, kul'tury, obrazovaniya i biznesa» (4–12 iyunya 2016, Sudak). M.: GPNTB Rossii, 2016. 3 p.

11. *Sochenkov I., Zubarev D., Tikhomirov I., Smirnov I., Shelmanov A., Suvorov R., Osipov G.* Exactus Like: Plagiarism Detection in Scientific Texts // Advances in Information Retrieval: 38th European Conference on IR Research, ECIR 2016 (March 20–23, 2016, Padua, Italy). Cham: Springer International Publishing, 2016. P. 837–840.

СВЕДЕНИЯ ОБ АВТОРАХ

ГУРЕЕВ Вадим Николаевич, кандидат педагогических наук, старший научный сотрудник информационно-аналитического центра Института нефтегазовой геологии и геофизики им. А.А. Трофимука СО РАН. Специалист в области библиометрии, владеет современными технологиями поиска, обработки и анализа библиографических данных. В сферу научных интересов входят библиометрический анализ отдельных научных направлений в России и за рубежом, оптимизация библиотечного комплектования, представление информации о публикационной активности организации в наукометрических базах данных, включающее проблемы идентификации метаданных.

Vadim Nikolaevich GUREYEV, Ph.D. (Education), senior researcher at Information Analysis Center of Trofimuk Institute of Petroleum Geology and Geophysics, SB RAS. Expert in bibliometrics, modern technologies of information retrieval, processing and analysis of bibliographic metadata. Area of expertise also includes bibliometric analysis of certain research areas in Russia and abroad, optimization of acquisition in libraries, representation of information on scholarly output in scientometric databases including problems of metadata identification.

GureyevVN@ipgg.sbras.ru

МАЗОВ Николай Алексеевич, кандидат технических наук, заведующий информационно-аналитическим центром Института нефтегазовой геологии и геофизики им. А.А. Трофимука СО РАН. Специалист в области информатики, библиографоведения, создания баз данных, а также наукометрии и библиометрии. В сферу научных интересов входят выявление трендов развития геологических наук, оценка показателей результативности научной деятельности, анализ публикационной активности российских научных организаций, библиометрия как отрасль библиотечного дела, анализ цитирования, определение плагиата, этика научных публикаций, экспертная оценка научных журналов.

Nikolay Alekseevich MAZOV, Ph.D. (engineering), head of Information Analysis Department of Trofimuk Institute of Petroleum Geology and Geophysics. Expert in library and information science, databases, as well as scientometrics and bibliometrics. His area of expertise comprises detection of research fronts in geosciences, evaluation of performance indicators of scientific work in Russia, bibliometrics as a branch of library science, citation analysis, plagiarism detection, publication ethics, expert evaluation of serials.

MazovNA@ipgg.sbras.ru

Материал поступил в редакцию 12 октября 2017 года