УДК 001.811+004.056

Н. А. Мазов, В. Н. Гуреев

Институт нефтегазовой геологии и геофизики им. А. А. Трофимука СО РАН,

ГПНТБ СО РАН

Результаты исследований по выявлению переводного плагиата с использованием библиометрических баз данных

В течение последних пяти лет авторы статьи исследуют возможность применения анализа цитирования к выявлению случаев переводного плагиата. За это время была теоретически обоснована работоспособность предложенного метода, а также проведён ряд прикладных исследований, позволивших выявить случаи неправомерных заимствований из текстов на других языках в нескольких научных областях. Показано, что метод особенно эффективен для естественных и точных наук в сравнении с гуманитарными, а также для анализа научных статей и отчётов в сравнении с другими типами публикаций.

Исследование выполнено при финансовой поддержке РФФИ в рамках научного проекта № 16-07-00652/17.

Ключевые слова: выявление плагиата, переводной плагиат, библиометрический анализ, анализ цитирования, библиометрические базы данных, Scopus, Web of Science, списки литературы.

UDC 001.811+004.056

Nikolay Mazov and Vadim Gureev

A. Trofimuk Institute of Petroleum Geology and Geophysics of the Russian Academy of Sciences Siberian Branch, Novosibirsk, Russia

State Public Scientific and Technological Library of the Russian Academy of Sciences Siberian Branch, Novosibirsk, Russia

Study results for the detection of translated plagiarism using bibliometric databases

In the last five years the authors have studied applicability of citation analysis to the problem of detection of translated plagiarism. During these years we verified the efficiency of this method and carried out several applied studies. It enabled us to detect a number of illegal uses of borrowed foreign texts in several research areas. We showed that the method is more efficient in natural and exact sciences as compared to social sciences. Besides, it is more appropriate when analyzing research articles, reviews and reports as compared to other document types.

Keywords: plagiarism detection, translated plagiarism, bibliometric analysis, citation analysis, bibliographic databases, Scopus, Web of Science, references.

One of the latest developments is the use of citation analysis for identification the cases of plagiarism. The proposed method is based not on a comparison of the text content, but on the comparison of the lists of the references of the two publications analyzed. Since references in publications of most natural and exact sciences are an obligatory element, and the number of references is, as a rule, sufficient for analysis, we within the framework of the proposed methodology one can directly address the content of texts, do not depend on lexical coincidences, as in most systems identification of plagiarism. This technology has been developed by a team of foreign researchers led by Gipp B. on the ground of the bibliographic coupling concept by Kessler M. In the bibliographic coupling for a unit of linkage between two articles, a common reference s are counted for two publications. Consequently, two articles are considered bibliographically related, and the strength of their bibliographic coupling, therefore, is the number of common references. The main characteristics of this method are its independence from the vocabulary and language of publications, as well as the possibility of automation. Therefore, it can directly apply also to the problem of identifying transfer plagiarism. In this case, a later work that has a similar literature list with an earlier publication, regardless of the language of the documents, becomes the object of analysis for possible borrowing. We propose to deal with to bibliographic sources instead of pretty expensive full-text processing. To implement our approach, we only need to subscribe to one or two multidisciplinary databases, for example, Web of Science or Scopus (access to which is now available in most Russian organizations as part of the national subscription) a study of publications with a suspiciously high number of coincidences, in particular, work with 41 general references, gave an additional result in the identification of several more instances of self-plagiarism in foreign publications, where the authors published their previous results, leaving the lists of literature practically unchanged.

Введение

Методы библиометрического анализа в последнее время находят всё большее применение – для выявления научных фронтов, поиска финансирующих организаций, коллабораторов и др. Одной из последних разработок, параллельно проводимой авторами этой работы [**1**, **2**] и коллективом зарубежных исследователей под руководством Б. Гиппа [**3**–**5**], является

использование такого библиометрического метода, как анализ цитирования, для выявления случаев переводного плагиата.

Следует отметить, что переводные заимствования пока не поддаются семантическому анализу, а поэтому не обнаруживаются существующими системами выявления плагиата, работающими только с каким-либо одним языком. Таким образом, на сегодняшний день переводной плагиат в наименьшей степени поддаётся обнаружению, и именно этим объясняется его популярность в научной среде неанглоязычных стран [**6**, **7**].

Предложенный метод основан не на сравнении текстового содержания, а на сличении пристатейных списков литературы (источников) двух анализируемых публикаций. Поскольку пристатейные списки в публикациях большинства естественных и точных наук являются обязательным элементом, а число ссылок, как правило, – достаточным для анализа, мы в рамках предложенной методики получаем возможность не касаться напрямую содержательной части текстов, не зависеть от лексических совпадений, как в большинстве систем выявления плагиата, а сличать только пристатейные списки литературы.

Обоснование метода анализа цитирования для выявления плагиата

Анализ цитирования при работе с пристатейными списками литературы двух публикаций основан на методе библиографического сочетания, или библиографической связи (*bibliographic coupling*), предложенном М. М. Кесслером [**8**, **9**]. При библиографическом сочетании за единицу связывания между двумя статьями принята общая ссылка из двух публикаций. Следовательно, 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 [Там же]. Основными характеристиками данного метода являются его независимость от лексики и языка публикаций, а также возможность автоматизации. Поэтому он напрямую может применяться и к выявлению переводного плагиата. В данном случае более поздняя публикация, имеющая схожий список литературы с более ранней, независимо от языка документов становится объектом анализа на наличие возможных заимствований.

Проблемы метода анализа цитирования для выявления плагиата

Идея сличения пристатейных списков литературы двух публикаций, включая анализ последовательности ссылок, выявление случаев масштабирования ссылок, анализ близости их расположения по отношению друг к другу в тексте и вероятность их совместного появления, – это оригинальное предложение группы исследователей, руководимой Б. Гиппом, и нашего коллектива в области библиометрического анализа.

К настоящему моменту группе Б. Гиппа удалось решить ряд технических проблем с использованием собственных программных разработок [**3**, **10**]. Определённым недостатком может выступать малое число ссылок, отчего мы признаём действенность метода лишь в предметных областях с достаточным числом пристатейных списков. Основной же проблемой остаётся поиск публикации, у которой был бы список литературы, схожий с подозрительной анализируемой работой.

Для поиска публикаций, являющихся возможным оригиналом, на основе которого создана вторичная публикация с неправомерными заимствованиями, необходим как можно более широкий доступ к полным текстам журналов различных издательств, отчётам различных фондов и пр. Системы выявления плагиата закупают доступ к архивам разных издательств, однако такой подход – дорогой, особенно в российских условиях. Оплатить подписку к коллекциям всех ведущих международных издательств на разных языках, на наш взгляд, трудновыполнимая задача.

Группа зарубежных исследователей под руководством Б. Гиппа ограничилась анализом баз данных открытого доступа, что позволило ей выявить достаточно большое число заимствований [**11**]. В то же время таких БД, несмотря на новые тенденции развития открытой науки и открытого доступа, пока немного, и в основном они ограничиваются биологической или медицинской тематикой.

Наш коллектив предложил обращаться не к полнотекстовым, а к библиографическим источникам.

Использование библиометрических баз данных для поиска источника плагиата

В отличие от обращения к полнотекстовым БД обращение только к библиографическим метаописаниям имеет два значительных преимущества.

Во-первых, предложенный подход дешев в использовании, поскольку нет требования к наличию полных текстов. В настоящее время во всех существующих программных решениях по выявлению заимствований в публикациях необходим доступ к полным текстам, поскольку сличаются именно тексты публикаций (см., напр., [**12**, **13**]). Для реализации нашего подхода

требуется подписка лишь на одну или две мультидисциплинарные БД, например *Web of Science* или *Scopus* (доступ к которым сейчас есть в большинстве российских организаций в рамках национальной подписки). Эти системы включают полное описание списков литературы любой индексируемой публикации и предоставляют инструменты поиска по пристатейной библиографии, чего вполне достаточно для обнаружения возможного оригинального источника.

Во-вторых, наш подход более эффективен, поскольку даже самая широкая подписка к полным текстам издательств не будет сопоставима с широтой покрытия научной литературы в мультидисциплинарных БД. На текущий момент и в *Web of Science*, и в *Scopus* насчитывается более 50 млн записей в каждой и более 1 млрд проиндексированных пристатейных ссылок. Охват журналов в *Web of Science* приближается к 18 тыс., а в *Scopus* – к 23 тыс. научных журналов.

Кроме того, учитывая наличие в базах данных веб-сервисов и интерфейсов *API*, мы получаем возможность автоматизировать всю технологическую цепочку – от создания поискового запроса на основе ссылок анализируемой публикации до получения из БД метаописаний публикаций со схожими списками литературы. Необходимо отметить, что технология поиска оригинала подозрительной публикации в полнотекстовых БД с использованием анализа цитирования также поддаётся автоматизации, что успешно доказала группа, руководимая Б. Гиппом.

Формирование поискового запроса
для обнаружения публикаций,
имеющих схожие списки литературы с анализируемой работой

Предложенный нами алгоритм обращения к библиометрическим БД состоит из следующих шагов:

1. Для каждого цитируемого источника из списка литературы подозрительной публикации формируется запрос в библиографическую БД с целью извлечь список публикаций, также цитировавших этот источник. Запрос может быть выполнен по таким метаданным источника, как авторы, заглавие, год публикации, название журнала или сборника, номер первой страницы, идентификатор *DOI* и др. Для сокращения числа выданных результатов и снижения объёма их последующей обработки возможно применение фильтрации по годам, при которой исключаются все годы, следующие за годом выхода подозрительной публикации (поскольку её первоисточник может быть опубликован только в тот же год или ранее). Работа проводится по идентификаторам публикаций – *eid* в *Scopus* или *ut* в *Web of Science*, что позволяет проводить формальную обработку записей в *Excel*.

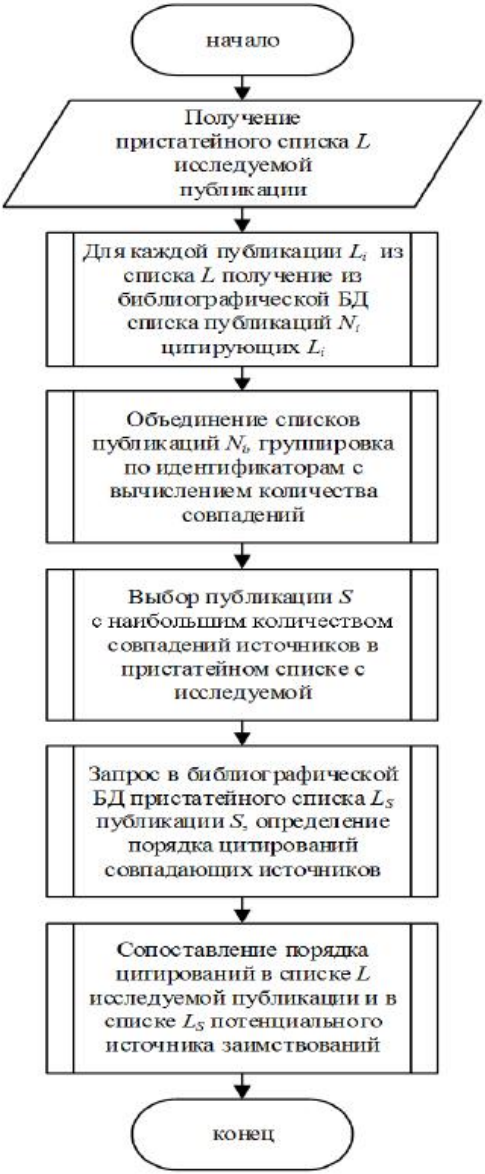

Схема формирования запроса и обработки его результатов с использованием библиометрической БД [1]

2. Далее мы проводим алфавитную сортировку выгруженных из БД идентификаторов источников, в которых цитировались те же публикации, что и в подозрительной работе. После этого проводится подсчёт числа совпадений.

3. Ранжированный по убыванию количества совпадающих источников список публикаций является предметом дальнейшего анализа на некорректные заимствования. В зависимости от обстоятельств исследования этот список может быть сокращён отсечением по абсолютной (например, количество совпадающих источников более 8) или относительной (например, количество совпадающих источников более 40% от всего списка литературы исследуемой статьи) границе.

4. В случае, если в исследуемой работе список литературы организован в порядке цитирования, формируется дополнительный запрос к библиометрической БД для получения списка литературы потенциального источника заимствований. Если этот список также организован в порядке цитирования, проводится анализ совпадения порядка цитирований.

Схематическое изображение алгоритма запроса в БД представлено на рисунке.

Выявленные случаи переводного плагиата с использованием анализа цитирований

Конкретными результатами исследований, проведённых нашей группой, стало выявление плагиата в различных типах научных публикаций. Были проверены несколько монографий, научных статей и обзоров, отчётов о НИР и диссертаций.

Метод тестировался в обеих мультидисциплинарных БД – *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 дал ссылки автор подозрительной публикации.

Нами были выявлены случаи плагиата в нескольких научных статьях, одной монографии, в главах которой – переводы из разных статей, а также в одном отчёте о НИР. Результаты анализа научных статей с использованием базы данных *Web of Science* представлены ранее в [**2**].

Тестирование метода на БД *Scopus* проводилось на примере отчёта о НИР в один из грантовых фондов. Этот случай представляет особенный интерес, поскольку был привлечён очень обширный ссылочный аппарат, что позволило во многом уточнить методику. Аналитический отчёт содержал 202 ссылки, 194 из которых были проиндексированы в *Scopus*. Отметим: в режиме расширенного поиска и в *Web of Science*, и в *Scopus* имеется возможность проводить глубинный поиск цитирующих статей даже для тех источников, которые не входят в основное индексируемое ядро.

На 194 работы из отчёта, которые были проиндексированы в *Scopus*, сделана 29 971 ссылка из 17 228 публикаций. Распределение ссылок оказалось следующим: по одному совпадению – в 12 711 публикациях, от 2 до 5 – в 4 440, от 6 до 10 – в 442, от 11 до 20 – в 108, от 21 до 30 совпадений – в 18, от 31 до 40 – в 6 публикациях. В трёх случаях было по 41, 65 и 82 совпадений. Анализ данных показал: отчёт был составлен из переводов трёх объёмных научных обзоров, где было по 82, 65 и 36 совпадений цитируемых источников соответственно.

Дальнейшее исследование публикаций с подозрительно высоким числом совпадений, в частности работы с 41 общей ссылкой, дало дополнительный результат – выявление ещё нескольких случаев самоплагиата в зарубежных публикациях, где авторы публиковали за небольшими разночтениями свои прежние результаты, оставляя списки литературы практически неизменными.

Проведённый нами анализ диссертационных работ, а также публикаций по общественным и гуманитарным наукам не дал положительных результатов, чему могут быть различные объяснения. Во-первых, обнаружилось очень скудное покрытие в международных БД цитируемых гуманитариями источников. Прежде всего это монографии, а также различные документы, неиндексируемые в этих БД, например, материалы газет, постановления, декларации и др.

Во-вторых, было отмечено намного более широкое – в сравнении с естественными и точными науками – время полужизни цитирований, уходящее вглубь XX в. Этот факт также накладывает ограничения на использование БД *Web of Science* и *Scopus*. В *Web of Science* глубина архива напрямую зависит от оплаченной подписки, и в нашем случае доступ был возможен только с 1975 г. Создатели же *Scopus* фокусируются прежде всего на широте охвата источников, а не на индексировании архивных материалов.

Заключение

За время исследований предложенный нами метод выявления переводного плагиата на основе анализа цитирований с использованием запросных строк в библиометрические БД показал положительные результаты. Автоматизация разработанной 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 этом направлении.

Промышленный запуск подобной системы в качестве дополнительных модулей в существующих системах выявления плагиата позволил бы, на наш взгляд, значительно повысить обнаружение случаев переводного плагиата и, как следствие, снизить объемы заимствований.

СПИСОК ИСТОЧНИКОВ

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. – DOI: 10.3103/S0147688216040092.

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. – DOI: 10.3103/S0147688213040151.

3. **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. – DOI: 10.1145/2484028.2484214.

4. **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.

5. **Gipp B., Meuschke N., Breitinger C., Pitman J., Nьrnberger 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. – V. 2, 2014. – P. 677–683. – DOI: 10.1109/JCDL.2014.6970168.

6. **Hvistendahl M.** China's Publication Bazaar // Science. – 2013. – V. 342, No 6162. – P. 1035–1039. – DOI: 10.1126/science.342.6162.1035.

7. **Мазов Н. А., Гуреев В. Н.** Публикации любой ценой? // Вестн. Рос. акад. наук. – 2015. – V. 85, № 7. – P. 627–631. – DOI: 10.7868/S0869587315050072.
Mazov N. A., Gureyev V. N. *Publikatsii lyuboy tsenoy? // Vestn. Ros. akad. nauk. – 2015. – V. 85, № 7. – P. 627–631. – DOI: 10.7868/S0869587315050072.*

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. – DOI: 10.1109/TIT.1963.1057800.

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. – DOI: 10.1002/asi.5090160309.

10. **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. – DOI: 10.1145/2034691.2034741.

11. **Meuschke N., Gipp B., Breitinger C.** CitePlag: A Citation-based Plagiarism Detection System Prototype // Proc. of the 5th International Plagiarism Conference (17–18 July, 2012, Newcastle upon Tyne, United Kingdom). – Edinburgh : iParadigms Europe Ltd, 2012. – P. 1–10.

12. **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 Междунар. конф. «Библиотеки и информационные ресурсы в современном мире науки, культуры, образования и бизнеса» (4–12 июня 2016 г., г. Судак). – Москва : ГПНТБ России, 2016. – С. 1–3.

Osipov G. S., Smirnov I. V., Tikhomirov I. A., Sochenkov I. V., Zuvarev D. V., Isakov V. A. *Tehnolodii semantiecheskogo poiska zaimstvovaniy v nauchnyih tekstah // Trudy 23-y Mezhdunar. konf. «Biblioteki i informatsionnye resursy v sovremennom mire nauki, kultury, obrazovaniya i biznesa» (4–12 iyunya 2016 g., g. Sudak). – Moskva : GPNTB Rossii, 2016. – C. 1–3.*

13. **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: 3^{8th} European Conference on IR Research, ECIR 2016 (March 20–23, 2016, Padua, Italy). – Cham : Springer International Publishing, 2016. – P. 837–840. – DOI: 10.1007/978-3-319-30671-1_76.

***Nikolay Mazov**, Cand. Sc. (Engineering), Head, Information and Library Center, A. Trofimuk Institute of Petroleum Geology and Geophysics of the Russian Academy of Sciences Siberian Branch, State Public Scientific and Technological Library of the Russian Academy of Sciences Siberian Branch;*

MazovNA@ipgg.sbras.ru

3, acad. Koptyug pr., 630090 Novosibirsk, Russia

***Vadim Gureev,** Cand. Sc. (Pedagogy), Leading Bibliographer, A. Trofimuk Institute of Petroleum Geology and Geophysics of the Russian Academy of Sciences Siberian Branch, State Public Scientific and Technological Library of the Russian Academy of Sciences Siberian Branch;*

GureyevVN@ipgg.sbras.ru

3, acad. Koptyug pr., 630090 Novosibirsk, Russia