К вопросу о разработке моделей выявления плагиата на основе цитирования с использованием наукометрических баз данных

On developing plagiarism identification models based on citation using scientometric databases

Н. А. Мазов, В. Н. Гуреев
Институт нефтегазовой геологии и геофизики
им. академика А. А. Трофимука СО РАН,
Новосибирск, Россия

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

Проблема переводного плагиата, при котором недобросовестные ученые переводят статьи с других языков и издают их на своем родном языке, является достаточно острой в научном мире. При этом такая форма плагиата является наиболее сложной для выявления. В докладе рассматривается относительно новый подход к обнаружению фактов неправомерного заимствования, основанный на поиске первоисточников с идентичными или схожими списками литературы. Данный подход успешно зарекомендовал себя в тестовых исследованиях группы зарубежных авторов, а также авторов настоящего доклада, и может быть рекомендован к промышленному внедрению в системы выявления плагиата в качестве дополнительного компонента.
Ключевые слова: плагиат, переводной плагиат, анализ цитирования, списки литературы, последовательность ссылок.

The problem of translated plagiarism when unprincipled researchers translate papers from foreign languages and publish them in their native language as their own works, is rather acute in the world of science. Moreover, this form of plagiarism is the hardest to detect. The authors propose a relatively new approach to improper borrowing detection based upon the search of primary sources with identical or similar lists of references. Such approach has proved itself in several testing studies, and those accomplished by the paper authors, too. This method can be recommended to be introduced into plagiarism detection systems as an add-on component
Keywords: plagiarism, translated plagiarism, citation analysis, bibliographies, reference sequence.

Введение

В настоящее время выявление плагиата в научных текстах, который в ответ на современные технологии его поиска приобретает все более закамуфлированные формы, стало задачей международного масштаба. Особенно остро проблема стоит в неанглоязычных странах, в том числе в Российской Федерации. Это связано как с определенной обособленностью от мирового научного сообщества, так и с тем, что все существующие разработки в сфере выявления интеллектуальных заимствований основаны на прямом сличении текстов и оказываются неэффективны, когда необходимо сравнить тексты на разных языках.

Определение подобных вторичных текстов сейчас возможно исключительно с привлечением экспертов в той или иной области. Данный подход имеет ряд существенных недостатков, основные среди которых – невозможность его массового применения, дороговизна и временные затраты. Таким образом, анализ работ на предмет заимствований данного типа не используется для проверки большинства научных статей, не применяется при проверке отчетов, которые пишутся в рамках грантов и государственных программ, не задействуется при проверке на оригинальность кандидатских и докторских диссертаций. Значительная для российской науки и ее рейтинга в мире проблема заключается в том, что в настоящих условиях недобросовестные ученые чувствуют себя в безопасности и зачастую отказываются от проведения самостоятельной и финансируемой государством работы в пользу менее трудозатратного перевода чужих публикаций. Следовательно,

количество оригинальных разработок, способных повысить авторитет отечественной науки, растет медленнее.

Как показал опыт с определением плагиата по типу копирования и вставки, эффективные результаты дает автоматизированный компьютерный анализ текстов. Данный подход мог бы дать значительные результаты и в выявлении обозначенных случаев плагиата, основанных на переводе текстов, с помощью автоматизированного сопоставления списков цитируемой литературы. Наш опыт одновременной работы в сфере перевода и научной библиографии позволил нам сделать наблюдение, что после перевода зарубежных публикаций авторы в своих статьях, как правило, сохраняют всю или значительную долю ссылок, также заимствованных из оригинальной статьи, и зачастую – их последовательность [1]. На основе этого можно с большой долей вероятности определить источник заимствования. До недавнего времени с небольшим опережением аналогичной разработкой занимался лишь один зарубежный коллектив, пришедший к аналогичным выводам и результатам [2-4].

Методы выявления плагиата

При сличении текста подозрительной публикации с оригинальным источником возможны четыре варианта, зависящие от стилей цитирования:

а) библиография и в оригинальном, и во вторичном тексте следует в порядке цитирования;
б) библиография и в оригинальном, и во вторичном тексте следует в алфавитном порядке;
в) библиография в оригинальном тексте следует в порядке цитирования, а во вторичном – в алфавитном порядке;
г) библиография в оригинальном тексте следует в алфавитном порядке, а во вторичном тексте – в порядке цитирования.

Для успешного решения задачи выявления плагиата, основанного на сличении моделей цитирования в двух текстах, мы использовали достижения библиометрии, а именно исследования, касающиеся библиографического сочетания (bibliographic coupling) [5] и социтирования [6, 7]. В библиометрии эти два метода используются при определении взаимосвязей между группами документов.

1. При библиографическом сочетании за единицу связывания между двумя статьями принята общая ссылка из двух публикаций. Таким образом, две статьи считаются библиографически связанными. Сила их библиографического сочетания, таким образом, – это количество общих для них ссылок.
2. Два документа считаются совместно процитированными, если они оба появились в списке ссылок третьего документа. Частота социтирования определяется как частота, с которой два документа цитируются вместе.

В настоящее время во всех существующих программных решениях по выявлению заимствований в публикациях необходим доступ к полным текстам, поскольку сличаются именно тексты публикаций. 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. По этому пути пошла группа упомянутых выше зарубежных исследователей.

На основе искусственно созданных примеров участниками зарубежного коллектива было создано три различных алгоритма моделирования цитирующих схем: сила библиографического сочетания (bibliographic coupling strength), наиболее длинная общая последовательность цитирований (longest common citation sequence) и жадное плиточное расположение цитирований (greedy citation tiling). За период с 2011 исследователи хорошо определили вероятность совместного появления одинаковых ссылок в двух текстах, которая зависит от времени создания цитируемых статей, их принадлежности различным дисциплинам, количества набранных ими цитирований (закон Матфея), тематики отраженных в публикациях исследований. Кроме того, исследователями были очерчены ограничения в применении метода, основное из которых – недостаточное количество ссылок. К прочим ограничениям относятся технические трудности непосредственного выявления моделей цитирования, которые могут сознательно внедряться плагиатором для сокрытия факта заимствования: перемешивание ссылок, их масштабирование, при котором та или иная

работа цитируется большее или меньшее количество раз в сравнении с оригинальным цитированием, вставка собственных цитат, иная разбивка текста, затрудняющая сличение количества слов, предложений, абзацев и параграфов между двумя ссылками. Указанной группой исследователей был создан прототип программы по выявлению случаев переводного плагиата (www.citeplag.org).

Основным недостатком в подходе, предложенном зарубежными исследователями, является зависимость от наличия полных текстов для анализа списков цитирований. Так, зарубежный коллектив в работе использовал полнотекстовую базу данных PubMed Central Open Access Subset. В условиях платного доступа к большинству коммерческих полнотестовых ресурсов такой подход во многом ограничивает фактологическую базу данных для сравнения, что особенно актуально для России. В предложенном нами подходе для выявления потенциального случая переводного плагиата необходимы лишь списки цитирований и доступ к мультидисциплинарным библиометрическим базам данных. Это существенно расширяет фактологическую базу данных для анализа, поскольку для сравнения оказываются доступны миллионы записей из библиографических баз данных.

К настоящему моменту нами была разработана и протестирована в ручном режиме технология создания поисковых запросов по полю пристатейной литературы в базе данных Scopus. Из списка пристатейной литературы анализируемой подозрительной статьи выгружался полный список пристатейной литературы, на основе элементов которого (автор – заглавие) строился единый сводный запрос в Scopus. Результаты запроса, выдающие публикации со схожими и – в редких случаях – идентичными списками пристатейной литературы, впоследствии анализировались в ручном режиме. По итогам анализа было выявлено несколько случаев масштабных заимствований из англоязычных текстов в ряде российских публикаций по биомедицинским наукам и информатике.

В Scopus в настоящее время отсутствуют развернутые эффективные поисковые возможности по спискам цитируемой литературы, однако в целом данная функция реализована и позволяет проводить масштабную обработку большого числа документов. Альтернативно имеется возможность использовать встроенный инструментарий и в базе данных Web of Science Core Collection. При внедрении возможности создавать поисковые запросы по спискам литературы в Российском индексе научного цитирования фактологическая база для исследований может существенно расшириться для поиска заимствований в отечественной научной литературе.

Заключение

Модель выявления плагиата на основе сличения в анализируемых публикациях списков цитирований и их последовательностей, на наш взгляд, работоспособна и ее тщательная разработка 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 и, как следствие, способствовать росту оригинальных отечественных исследований.

Благодарности. Разработка модели выявления плагиата в 2016 г. получила поддержку РФФИ и выполняется в рамках научного проекта № 16-07-00652\16.

Литература

1. Гуреев В. Н., Мазов Н. А. Анализ цитирования как основа для разработки дополнительного модуля в системах антиплагиата // Научно-техническая информация. Серия 1: Организация и методика информационной работы. – 2013. – Т. 12. – С. 12–15.
2. Gipp B., Meuschke N., Breitinger C., Lipinski M., Nürnberger A. Demonstration of citation pattern analysis for plagiarism detection // Book Demonstration of citation pattern analysis for plagiarism detection / Editor. – Dublin, 2013. – P. 1119–1120.
3. Gipp B., Meuschke N., Breitinger C. Citation-based plagiarism detection: Practicability on a large-scale scientific corpus // Journal of the Association for Information Science and Technology. – 2014. – V. 65, № 8. – P. 1527–1540.

4. Gipp B., Meuschke N. Citation Pattern Matching Algorithms for Citation-based Plagiarism Detection: Greedy Citation Tiling, Citation Chunking and Longest Common Citation Sequence // Book Citation Pattern Matching Algorithms for Citation-based Plagiarism Detection: Greedy Citation Tiling, Citation Chunking and Longest Common Citation Sequence / Editor. – Mountain, View, CA, USA: ACM, 2011. – P. 1–10.
5. Kessler M. M. An Experimental Study of Bibliographic Coupling Between Technical Papers // IEEE Transactions on Information Theory. – 1963. – V. 9, № 1. – P. 49–51.
6. Small H. Co-citation in the scientific literature: a new measure of the relationship between two documents // Journal of the American Society for Information Science. – 1973. – V. 24, № 4. – P. 265–269.
7. Маршакова И. В. Система связей между документами, построенная на основе ссылок: по данным Science Citation Index // Научно-техническая информация. Серия 2: Информационные процессы и системы. – 1973. № 6. – С. 3–8.