Мазов Николай Алексеевич, к.т.н., заведующий информационно-аналитическим центром Института нефтегазовой геологии и геофизики им. А.А. Трофимука СО РАН

Гуреев Вадим Николаевич, к.п.н., научный сотрудник информационно-аналитического центра Института нефтегазовой геологии и геофизики им. А.А. Трофимука СО РАН

Косяков Денис Викторович, заместитель директора по информационно-коммуникационным технологиям, научный сотрудник информационно-аналитического центра Института нефтегазовой геологии и геофизики им. А.А. Трофимука СО РАН

**Проблемы автоматизации процессов выявления плагиата
на основе анализа цитирования**

Метод анализа цитирований в научных публикациях имеет множество различных практических приложений. Одной из последних разработок, параллельно проводимой коллективом зарубежных исследователей под руководством Б. Гиппа и авторами настоящей работы, 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. Таким образом, метод позволяет выявлять такие трудные для определения формы плагиата, как плагиат с существенным перефразированием, переводной плагиат или плагиат идей.

Вслед за завершенным теоретическим обоснованием возможности использования данного подхода в качестве дополнительного модуля в системах выявления плагиата встала задача автоматизации процессов поиска возможных оригиналов при анализе подозрительных публикаций. Группа зарубежных исследователей во главе с Б. Гиппом разработала прототип автоматизированной системы сопоставления ссылок в анализируемых публикациях на основе анализа полных текстов (http://www.citeplag.org). С учетом ограниченности подобного подхода, связанного с дороговизной доступа к полнотекстовым базам данных, их распределенностью по разным издательствам и частым отсутствием в них интерфейса программирования приложений API, нами был предложена концепция использования мультидисциплинарных библиографических баз данных для анализа цитирования с целью выявления плагиата. В качестве библиографической базы данных может использоваться любая система с поддержкой возможности для просмотра списков цитировавших публикацию работ.

На примере базы данных Scopus, выбор которой обусловлен политематическим охватом источников и наиболее широким диапазоном покрытия современной научной литературы, мы реализовали два подхода к извлечению списков публикаций, цитирующих тот же источник, что и исследуемая работа. Первый подход использует возможность формирования запроса по пристатейным спискам (Cited Reference Search) с использованием инструментария в Scopus. Аналогичные запросы позволяет также выполнять интерфейс баз данных Web of Science. Второй подход предполагает поиск источника в базе данных с последующим извлечением списка цитировавших этот источник публикаций. Этот режим можно использовать также при работе с Российским индексом научного цитирования.

Преимуществом первого подхода является возможность нахождения списков цитирующих публикаций для источников, непроиндексированных в системе, второй подход позволяет снизить ошибки при нахождении в базе данных источника.