Технологии семантического поиска заимствований в научных текстах

Technologies for semantic plagiarism detection in scientific texts

Г. С. Осипов, И. В. Смирнов, И. А. Тихомиров, И. В. Соченков, Д. В. Зубарев, В. А. Исаков
Институт системного анализа Федерального исследовательского центра «Информатика и управление» РАН,
Москва, Россия

Gennady Osipov, Ivan Smirnov, Ilya Tikhomirov, Ilya Sochenkov, Denis Zubarev and Vadim Isakov
Institute for Systems Analysis of FRC CSC, Russian Academy of Sciences,
Moscow, Russia

В докладе представлены технологии семантического поиска заимствований, предназначенная для выявления смысловых текстовых заимствований в научных публикациях. Рассмотрены принципы работы и функциональные возможности.

The paper describes technologies for semantic plagiarism detection in scientific publications. The basic functionality and features are described.

Технологии семантического поиска текстовых заимствований, разработанные в Институте системного анализа ФИЦ ИУ РАН, основаны на методах реляционно-ситуационного анализа текстов [1]. В отличие от других аналогичных технологий и систем поиска заимствований, наши технологии используют глубокую лингвистическую обработку текстов, включая морфологический, синтаксический и семантический анализ [2]. Использование результатов лингвистического анализа текстов позволяет выявлять не только дословные заимствования, но и заимствования с перестановками слов и предложений местами, заменой слов и словосочетаний синонимами, разбиением и объединением предложений. В таблице 1 приведены примеры выявляемых заимствований со значительным перефразированием и их источников. Жирным выделены совпадающие слова.

Таблица 1 – Примеры выявляемых перефразированных заимствований

Проверяемый текст	**Текст источника**
Текст докладной записки делится на две части: Констатирующая (описательная), где излагаются имевшие место факты или описывается ситуация, вторая, где излагаются предложения, просьбы.	**Докладная записка** обычно состоит из **двух частей:** в первой **описывается** сложившаяся **ситуация,** во **второй излагаются предложения, просьбы,** делаются конкретные выводы.
Сам метод заключается в следующем: на каждом шаге мы выбираем один из исходных элементов и вставляем его на нужную позицию в уже отсортированном списке, до тех пор, пока набор исходных данных не будет исчерпан.	На **каждом шаге** алгоритма **мы выбираем один** из **элементов** входных **данных** и восстанавливаем **его** на **нужную позицию** в **уже отсортированном списке,** до **тех пор пока набор** входных **данных не будет исчерпан**.

Технологии семантического поиска текстовых заимствований работают с текстами на русском и английском языках. Алгоритмы семантического поиска заимствований прошли независимую объективную проверку на международных соревнованиях по поиску заимствований CLEF-2014 и показали высокие результаты по качеству и скорости поиска заимствований [3].

Технологии семантического поиска текстовых заимствований предоставляют следующие возможности:

- обработка документов в любом распространенном текстовом формате (DOC, DOCX, PDF, TXT и т.д.);
- выявление в проверяемом тексте смысловых заимствований из других текстов. Для каждого найденного источника заимствований отображаются заимствованные фрагменты и вычисляется процент заимствованных из него фрагментов;
- определение условно корректных заимствований – заимствований из источников, на которые есть ссылки в проверяемом документе.
- оценка оригинальности текста, отражающей степень уникальность текста с учетом выявленных заимствованных фрагментов. Степень оригинальности документа определяется как процент заимствованных фрагментов от общего объёма проверяемого текста;
- учёт общеизвестных фрагментов. Общеизвестными считаются фрагменты, встречающиеся в большом количестве документов информационной базы системы. Такие фрагменты не считаются заимствованиями и не учитываются при определении степени оригинальности документа.

Технологии семантического поиска текстовых заимствований предоставляют удобные средства просмотра и оценки корректности заимствований:

- возможность отобразить проверяемый документ в исходном формате;
- удобная навигация по страницам документа;
- подсветка заимствований цветами, соответствующими источникам заимствований;
- отключение источников заимствований и отдельных фрагментов из рассмотрения;
- автоматический пересчет оригинальности.

На рисунке 1 представлена область просмотра и редактирования текстовых заимствований.

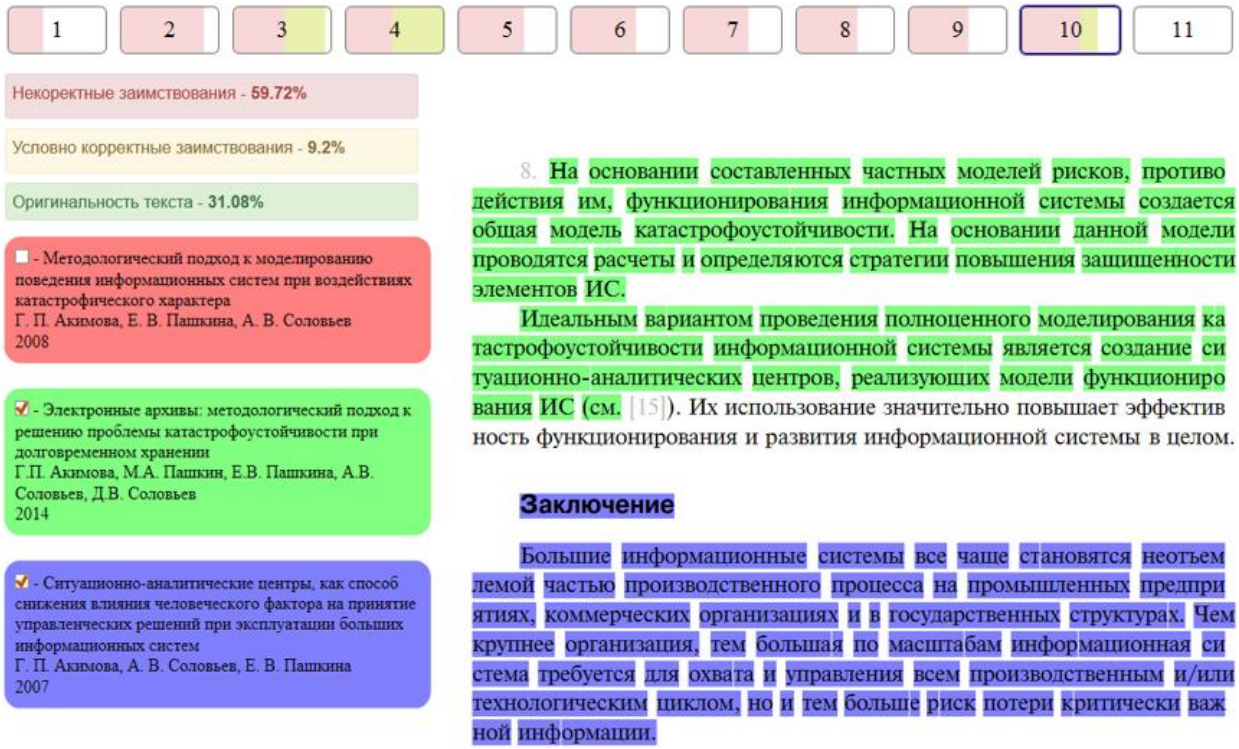

Рис. 1. Область просмотра заимствований

Демонстрация возможностей технологий семантического поиска заимствований доступна в системе Exactus Like по адресу http://like.exactus.ru.

Технологии семантического поиска текстовых заимствований интегрированы в программно-аппаратный комплекс интеллектуального поиска и анализа больших массивов текстов TextAppliance – промышленном решении, предназначенном для комплексной обработки больших коллекций текстов (http://textapp.ru/). TextAppliance состоит из сервера (или группы серверов, объединенных в кластер) и интеллектуальных программных сервисов поиска и анализа больших коллекций текстовых документов. TextAppliance с сервисами семантического поиска текстовых заимствований легко интегрируется в любую информационную инфраструктуру за счёт унифицированных программных интерфейсов и может быть быстро развёрнут в любой организации.

На основе TextAppliance реализован ряд систем обработки научных текстов, в их числе система поиска текстовых заимствований РУКОНТЕКСТ (http://text.rucont.ru/), ориентированная в том числе на проверку студенческих работ в ВУЗах.

Литература

1. Осипов Г.С., Смирнов И.В., Тихомиров И.А. Реляционно-ситуационный метод поиска и анализа текстов и его приложения // Искусственный интеллект и принятие решений – №2. – 2008. – С. 3-10.
2. Ilya Sochenkov, Denis Zubarev, Ilya Tikhomirov, Ivan Smirnov, Artem Shelmanov, Roman Suvorov and Gennady Osipov. "Exactus Like: Plagiarism Detection in Scientific Texts." In Advances in Information Retrieval, pp. 837-840. Springer International Publishing, 2016.
3. D. Zubarev, Sochenkov, I. Using Sentence Similarity Measure for Plagiarism Source Retrieval – Notebook for PAN at CLEF 2014. In: CEUR Workshop Proceedings, CEUR-WS.org, Eds. L. Cappellato, N. Ferro, M. Halvey and W. Kraaij. 2014. P.p. 1027–1034, / [Электронный ресурс] URL: http://ceur-ws.org/Vol-1180/CLEF2014wn-Pan-ZubarevEt2014.pdf, (дата обращения 27.04.2015).