Exactus Like – система выявления заимствований в научных текстах

Exactus Like – system for detecting reused text in scientific publications

Г. С. Осипов, И. В. Смирнов,
И. А. Тихомиров, И. В. Соченков, Д. В. Зубарев
Институт системного анализа РАН,
Москва, Россия

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

В докладе представлена система Exactus Like, предназначенная для выявления смысловых текстовых заимствований в научных публикациях. Рассмотрены принципы работы и функциональные возможности системы.

The paper presents Exactus Like – the system for detecting reused text in scientific publications. The basic functionality and features of the system are considered.

Система Exactus Like предназначена для выявления смысловых заимствований в научных текстах и оценки оригинальности научных текстов. Система является инструментом, помогающим студентам, аспирантам, научным сотрудникам в написании рефератов, курсовых, дипломных работ, кандидатских диссертаций и авторефератов, научных статей и других научных работ. Система также может быть использована преподавателями для проверки студенческих работ, редакторами научных журналов и издательств, организаторами научных конференций для проверки поступающих научных публикаций на оригинальность. Кроме того, система может быть полезна диссертационным советам для проверки диссертаций и различным ведомствам для проверки оригинальности отчётов о НИР.

Метод выявления смысловых заимствований

Подходы к анализу научных текстов, используемые в системе Exactus Like, основаны на методах реляционно-ситуационного анализа текстов [1]. В отличие от других аналогичных систем, в Exactus Like тексты подвергаются глубокому лингвистическому анализу, включая морфологический и синтактико-семантический анализ [2]. Использование результатов лингвистического анализа текстов позволяет выявлять не только дословные заимствования, но и смысловые заимствования с учетом перефразирования, замены слов синонимами, перестановки местами слов и предложений. В таблице 1 приведены примеры выявленных системой заимствований со значительным перефразированием и их источников. Жирным выделены совпадающие слова.

Таблица 1 – Примеры выявленных перефразированных заимствований

Проверяемый текст	**Текст источника**
В безглагольных предложениях или предложениях, для которых предикатное слово не найдено в словаре, синтаксемы присутствуют рядом с другими элементами предложения, и несут свое значение только в данном контексте.	В безглагольных предложениях синтаксемы присутствуют рядом с другими элементами предложения и несут своё значение только в данном контексте.
Прямые прокси содержат адрес сервера, т.е. имя хоста и номер порта.	Прямые прокси (direct proxy) содержат информацию об адресе сервера в виде идентификатора протокола, имени хоста и номера порта

Метод выявления смысловых заимствований работает с текстами на русском и английском языках. Алгоритмы Exactus Like прошли независимую объективную проверку на международных соревнованиях по поиску заимствований CLEF-2014 и показали высокие результаты по качеству и скорости поиска заимствований [3].

Функциональные возможности системы

Система Exactus Like предоставляет следующие возможности:

- автоматическое формирование информационной базы системы – больших коллекций научных публикаций из различных источников, включая интернет. Тексты научных публикаций подвергаются лингвистическому анализу, создаются семантические индексы. При формировании коллекций выполняется автоматическое извлечение из научных публикаций метаинформации – названия, авторов и года публикации. Полученная метаинформация используется для дальнейшего анализа;
- загрузка пользователем на сервер научной публикации в виде файла в любом распространенном текстовом формате (DOC, DOCX, PDF, TXT и т.д.) или в виде текста, вводимого в форму. Для загружаемой публикации можно указать год опубликования, который будет учитываться при оценке оригинальности. Загружаемая публикация в системе не сохраняется;
- выявление в загруженном документе смысловых заимствований из других текстов, хранящихся в базе системы. Для каждого найденного источника заимствования отображаются заимствованные фрагменты и вычисляется процент заимствованных из него фрагментов. Документ считается источником заимствования, если дата его публикации не позднее даты анализируемого текста;
- поиск в информационной базе системы документов, содержащих заимствования из загруженной публикации. Документ считается заимствующим, если дата его публикации позднее даты публикации анализируемого текста. Для каждого найденного заимствующего документа отображаются фрагменты, заимствованные из загруженного текста, и процент таких заимствований от общего числа проверяемых фрагментов;
- определение условно корректных заимствований – заимствований из источников, на которые есть ссылки в анализируемом документе.
- оценка оригинальности загруженной публикации, отражающей степень уникальность текста с учетом выявленных заимствованных фрагментов. Степень оригинальности документа определяется как процент заимствованных фрагментов от общего числа проверяемых фрагментов загруженного 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;
- определение значимых оригинальных фрагментов. Используя эту функцию, пользователь может просмотреть наиболее значимые фрагменты загруженного текста, для которых не найдено заимствований;
- поиск заимствований в интернет. Если пользователя не удовлетворяют результаты поиска заимствований в информационной базе системы, то можно воспользоваться функцией поиска заимствований в сети интернет с помощью известных поисковых машин (Yandex, Google и т.д.).

Информационная база системы

На данный момент информационная база системы насчитывает 8 млн. 184 тыс. научных документов и включает следующие коллекции:

- студенческие рефераты (303 тыс. документов);
- материалы некоторых российских и зарубежных конференций (35 тыс. документов);
- авторефераты кандидатских и докторских диссертаций (45 тыс. документов);

- публикации российских рецензируемых журналов, включая публикации из Киберленинки и MathNet (586 тыс. документов);
- российские энциклопедии (512 тыс. статей);
- публикации на английском языке с ресурса Arxive.org (762 тыс. документов);
- англоязычная Википедия (4 млн. 759 тыс. статей);
- русскоязычная Википедия (1млн. 178 тыс. статей).

Информационная база постоянно пополняется новыми документами.

Использование системы

Система Exactus Like является частью системы и технологий интеллектуального поиска и анализа научных публикаций Exactus Expert, предназначенной для информационно-аналитической поддержки научной деятельности [4]. Демонстрационная версия системы доступна по адресу http://like.exactus.ru.

Система функционирует в распределенной вычислительной среде на кластерной установке. Система ориентирована на обработку больших массивов данных и легко масштабируется. Exactus Like интегрируется в любую информационную инфраструктуру за счёт унифицированных программных интерфейсов и может быть быстро развёрнута в любой организации.

В дальнейшем планируется пополнение информационной базы новыми документами, ускорение методов выявления заимствований и оценки оригинальности научных публикаций, расширение типов отчётов о найденных заимствованиях.

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. Shelmanov A. O., Smirnov I. V., Methods for Semantic Role Labeling of Russian Texts // Computational Linguistics and Intellectual Technologies: Papers from the Annual International Conference «Dialogue» (2014). Issue 13 (20). – 2014. – pp. 607-619.
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).
4. Osipov, G., Smirnov, I, Tikhomirov, I., Sochenkov, I., Shelmanov, A., and Shvets, A. Information Retrieval for R&D Support // Paltoglou, Georgios, Loizides, Fernando, Hansen, Preben (Eds.) Professional Search in the Modern World. – Lecture Notes in Computer Science (LNCS). – Springer International Publishing, 2014. – Vol. 8830 – pp. 45–69.