УДК 519.24

Определение заимствований в тексте без указания источника

К.Ф.Сафин[1,2], М.П.Кузнецов[1], Р.В.Кузнецова[1,2], В.В.Стрижов[1]
[1]Московский физико-технический институт (государственный университет)
[2]ЗАО “Анти-плагиат”

Текстовый плагиат является большой проблемой в сфере образования и научных исследований. Развитие сети интернет сделало большой объем информации свободным для плагиата.

В задаче обнаружения плагиата существует два глобальных подхода: выявление “внешнего” и “внутреннего” плагиата. При поиске внешних заимствований есть сторонний корпус, из которого возможны заимствования. То есть задача состоит в попарном сравнении участков подозрительного текста и текстов из корпуса заимствований и нахождении сходства. Стандартная задача поиска внутреннего плагиата [1,2,3] ставится следующим образом: для заданного документа необходимо определить, написан он полностью одним автором или содержит фрагменты заимствований. При этом предполагается, что у текста есть один главный автор, которым написано не менее 70% текста. В отличие от обнаружения “внешнего” плагиата, задача поиска “внутреннего” плагиата должна решаться без привлечения внешнего корпуса документов. Алгоритмы детектирования внутреннего плагиата должны анализировать стиль письма, т.е. выделять характерные признаки, свойственные данному автору.

В данной работе исследуется модификация алгоритма [1], являющегося победителем конкурса PAN-2011 по выявлению внутреннего плагиата. PAN – международный конкурс, который проводится с 2007 года с целью нахождения новых алгоритмов поиска текстовых заимствований. Алгоритм строит статистическое описание текста, которое используется для выявления заимствованных фрагментов. Статистика должна удовлетворять следующим условиям: на оригинальных фрагментах текста она должна иметь небольшой разброс значений, а на заимствованных иметь значительные отличия.

Предлагаемый алгоритм состоит из трех основных этапов:

1. Деление текста на элементарные сегменты, среди которых будут искаться заимствованные
2. Построение статистики для каждого сегмента, основанной на частоте встречаемости слов
3. Анализ значений ряда статистики на выбросы

Алгоритм был настроен и протестирован на полном корпусе документов PAN-2011. Качество работы сравнимо с алгоритмом, взятым за основу данной работы.

Литература

1. *Oberreuter G., L'Huillier G., Rios S., Velasquez J.* Approaches for intrinsic and external plagiarism detection. 2011.
2. *Stamatatos E.* Intrinsic plagiarism detection using character n-gram profiles. 2009.
3. *Potthast M., Stein B., Barron-Cedeno A., Rosso P.* An evaluation framework for plagiarism detection. 2011.