УДК 347.214.23:004.021

Студ. Д.Д. Курмашев
Науч. рук. ст. преп. Е. А. Блинова
(кафедра информационных систем и технологий, БГТУ)

АЛГОРИТМЫ И МЕТОДЫ ВЫЯВЛЕНИЯ УНИКАЛЬНОСТИ ТЕКСТОВЫХ ДОКУМЕНТОВ

Одной из наиболее значимых возможностей, появившихся в руках пользователей после глобального распространения сети Интернет, стала возможность копировать и распространять информацию.

Существует ряд методов, которые характеризуются по типу оценки сходства. Глобальная оценка использует большие части текста или документа для нахождения сходства в целом, в то время как локальные методы на входе проверяют ограниченный сегмент текста. В настоящее время наиболее распространенным подходом является метод дактилоскопии, суть которого заключается в выборе из ряда документов набора из нескольких подстрок, которые и являются «отпечатками». Рассматриваемый документ будет сравниваться с «отпечатками» для всех документов коллекции. Найденные соответствия с другими документами указывают на общие сегменты текста. Проверка документа дословным перекрытием текста представляет собой классическое сравнение строк.

Также существуют виды анализа на основе последовательностей частей речи. В качестве параметров разбиения берутся различные последовательности частей речи. И в результате для текста находятся последовательности, которые выделяли из текстов фрагменты, то есть алгоритм выделяет из текста фрагменты неоднородности, имеющие разные частоты встречаемости выбранной последовательности частей речи, что показывает на возможный плагиат в данном месте.

На данный момент в открытых источниках можно найти ссылки на ряд программных средств, которые реализуют данные методы и алгоритмы анализа информации. Но стоит отметить, что проблема не решена, так как достоверно определить является ли информация копией с другого ресурса или нет нельзя, потому что теоретически возможен вариант, когда несколько людей начнут разработку схожих материалов примерно в одно и тоже время. Мировая история имеет ряд примеров подобных случаев.