ТЕХНИЧЕСКИЕ НАУКИ

Зиберт Андрей Оскарович
магистрант
Хрусталев Виталий Игоревич
канд. техн. наук, доцент

ФГБОУ ВПО «Хакасский государственный
университет им.Н.Ф. Катанова»
г. Абакан, Республика Хакасия

СОВРЕМЕННЫЕ МЕТОДЫ ОПРЕДЕЛЕНИЯ НАЛИЧИЯ ЗАИМСТВОВАНИЙ В ТЕКСТЕ

Аннотация*: в статье описывается процесс разработки алгоритма определения наличия заимствований в тексте с использованием эталонного множества слов. Рассматриваются достоинства и недостатки данного алгоритма, а также приводится описания различных методов создания эталонного множества слов.*

Ключевые слова*: плагиат, заимствование, метод шинглов, метод сравнения, часто встречающиеся слова.*

В настоящее время проблема выявления плагиата становится все более актуальной. Даже в работах, защищаемых в высшей аттестационной комиссии, выявляются факты плагиата и некорректного заимствования в текстах научных работах и научно – исследовательских работ, а санкции, связанные с этими фактами, распространяются не только на авторов работ, но и на их руководителей и оппонентов [1].

Ранее нами было произведено исследование и установлен наиболее оптимальный алгоритм с точки зрения затрат времени и результата обработки текстов алгоритм сравнения двух текстов и определения наличия заимствований в одном

тексте относительного другого текста [2, с. 37]. Им стал алгоритм, базирующийся на комбинации метода шинглов и метода сравнения наиболее часто встречающихся слов.

Суть метода состояла в создании для каждого исследуемого текста подмножества, состоящего из 30% наиболее встречающихся слов, упорядочивании данного подмножества по алфавиту и сравнения полученных подмножеств с использованием метода шинглов.

Но в дальнейшем был выявлен недостаток метода, связанный с тем, что авторы при написании текста работы будут иметь доступ как своей работы, так и к оригинальной работе и смогут изменять текст своей работы до тех пор, пока процент пересечений шинглов не снизится до определенного уровня и, таким образом, система, базирующаяся на данном алгоритме, получит некорректный результат.

В связи с этим был модифицирован алгоритм сравнения двух текстов, который более не базируется на сравнении шинглов, а работающий на оценке некоторой числовой характеристике сравниваемых текстов. Для этого было введено так называемое эталонное множество слов и производилось сравнение пересечений множеств слов двух исследуемых текстов с эталонным множеством. Суть работы алгоритма отражена на рисунке 1.

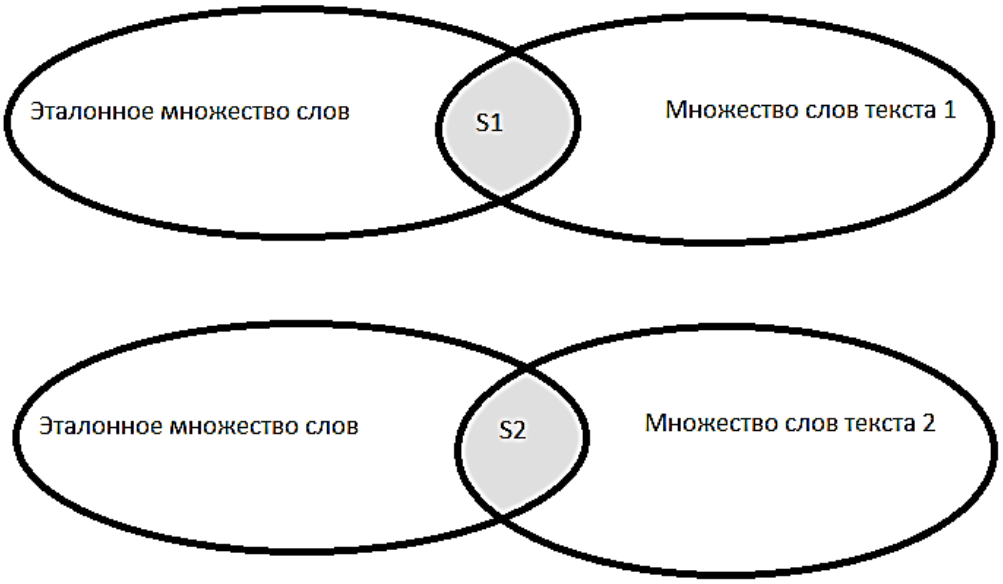

Рис. 1. Иллюстрация работы алгоритма

Одним из ключевых элементов правильной работы алгоритмов в данном случае является составление эталонного множества слов. В ходе разработки модифицированного алгоритма было предложено три способа генерации эталонного множества:

1. Статичное множество слов, созданное на основании большого числа текстов научных и научно-исследовательских работ по нескольким направлениям.

2. Статичное множество слов, созданное на основе текстов по направлению, совпадающему с направлением исследуемых текстов.

3. Динамическое формирование эталонного множества для каждого пары исследуемых текстов.

Далее для оценки процента заимствований было произведено попарное сравнение оригинальных текст и текстов с наличием элементов заимствований. Кроме того, для оценки корректности работы алгоритма с текстами, не содержащих элементов заимствований была произведена попарная проверка первого оригинального текста с остальными оригинальными текстами. Количество пересечений множеств слов исследуемых текстов с эталонными приведена в таблице 1.

Таблица 1

Количество пересечений множеств слов

Номер текста	Статичное множество слов на основе работ по всем направлениями	Статичное множество слов на основе текстов по направлению информатика	Динамически формируемое множество слов
1	24	28	32
2	37	36	44
3	8	8	15
4	42	44	47
5	13	12	13

Как и было ожидаемо, статичные методы формирования множеств слов показали худший результат, чем динамически формируемое множество слов, иногда даже давая неправильный результат (как в случае сравнения 1 и 3 текстов, а также 1 и 5 текста). Данный факт объясняется тем, что в таких множествах не учитывается специфика текста, так как, например, в обзорных статьях, не относящихся к узкоспециализированной тематике встречается много общенаучных

терминов, употребление которых в работах не является фактами наличия заимствований. Динамически формируемое множество слов позволяет избегать данного явления.

Список литературы

1. Из экспертных советов ВАК исключили руководителей и оппонентов плагиаторов…/ [Электронный ресурс]. – Режим доступа: http://lenta.ru/news/2014/11/21/auzan/ (дата обращения: 02.02.2015).

2. Зиберт А.О., Хрусталев В.И. Разработка системы определения наличия заимствований в работах студентов высших учебных заведений. Алгоритмы поиска нечетких дубликатов // Universum: Технические науки: электрон. научн. журн. – 2014. – №3 (4) [Электронный ресурс]. – Режим доступа: http://7universum.com/ru/tech/archive/item/1139 (дата обращения: 02.02.2012).