С. О. Жихарев, Н. Б. Осипенко

(ГГУ им. Ф. Скорины, Гомель)

РАЗРАБОТКА ПРОГРАММНЫХ СРЕДСТВ ПРОВЕРКИ УНИКАЛЬНОСТИ ТЕКСТА С ИСПОЛЬЗОВАНИЕМ РЕСУРСОВ ПОИСКОВЫХ СИСТЕМ

В работе описываются программные средства проверки уникальности контента (content – содержание, суть) с использованием ресурсов поисковых систем. Контент – это то, что люди ищут в Интернете: тексты, видео, изображения, аудио и т.д. В большинстве случаев под контентом принято понимать именно текстовую информацию. Уникальный контент необходим каждому Интернет-ресурсу по двум причинам. Во-первых, уникальный контент – это залог высокой посещаемости ресурса. Так, если посетитель зайдет на сайт и увидит информацию, которую он раньше нигде не видел и не находил, он заинтересуется этим сайтом и обязательно вернется, надеясь получить снова уникальный контент. Второй фактор – поисковый, или механический. Здесь уникальный контент необходим для достижения высоких позиций в результатах поисковой выдачи.

Уникальность текста проверяется на основе метода шинглов (shingl –крыть щепой), смысл работы которого заключается в том, чтобы проанализировать разбитые на кусочки цепочки слов проверяемых текстов и сравнить полученные результаты. Чем короче отрезки текста, тем точнее выходит результат.

Текст разбивается на отрезки из пяти-шести слов, которые преобразуются в двоичный код. Во время проверки каждому участку текста присваивается некое уникальное число, и результат вычисляется по заданному алгоритму в виде контрольной суммы. У разных текстов контрольные числа не могут совпадать. Причем в процессе подобного «сканирования» текста конечное слово одного отрезка является первым для нового отрезка, и, следовательно, ни одно слово не теряется.

Общая схема действия проверки текста на уникальность состоит из трех этапов:

1. Канонизация текста – процесс исключения из текста всех слов и символов, не имеющих смысловой нагрузки;
2. Разбиение канонизированного текста на шинглы;
3. Сравнение шинглов одного текста с шинглами другого.

Разработанное в ходе работы приложение можно позиционировать как законченный продукт, позволяющий производить проверку уникальности текстового контента в сети Интернет. Разработка приложения проводилась в среде программирования Microsoft Visual C# 2010 Express.