Для некоторых это врожденное качество, но другие должны работать над этим.

Можно говорить и об очевидных качествах: любви к работе, самоотверженности, искренности, страсти к обучению, хорошей трудовой этике и т. д., но они являются универсальными и должны быть присуще любому человеку, который устраивается на работу.

В области тестирования существует множество мифов о профессии тестировщика. Один из наиболее серьезных мифов заключается в том, что тестировщики отвечают за качество. Это не так: за качество отвечают абсолютно все, кто работает над продуктом. Тестировщики не управляют исходным кодом, каталогами, областью действия продукта, бюджетом, наймом-увольнением ответственных лиц, контрактами с клиентами и так далее. Они отвечают за выявление потенциальных угроз для качества.

Таким образом, основные задачи, стоящие перед профессионалами в области тестирования – моделировать различные, в том числе форс-мажорные ситуации, выявляя дефекты в программных системах, прогнозировать вероятные сбои, соотносить конечные результаты с начальными планами. При этом тестировщик должен обладать определенными личностными качествами, главным из которых, на мой взгляд, является любопытство.

УДК 621.762.4

Ковалевский А. Н.

АЛГОРИТМ ШИНГЛОВ

БНТУ, г. Минск
Научный руководитель: канд. техн. наук, доцент Дробыш А. А.

Алгоритм шинглов (от английского shingles – гонт или дранка) – алгоритм, разработанный для нечеткого поиска дубликатов текста. Нечеткий поиск дубликатов означает что в

дубликате возможно, как копирование всей строки, так и только отдельных словосочетаний. Данный алгоритм используется для поиска копий и дубликатов текста, например, в многочисленных онлайн сервисах по проверке уникальности текстов хотя в последнее время некоторые самые распространенные переходят на другие алгоритмы собственной разработки, а также используется поисковыми системами для противодействия поисковому спаму, исключая из результата поиска идентичные тексты.

Сам алгоритм состоит из пяти этапов: канонизация текста, разбиение на шинглы, вычисление хэшей шинглов, случайная выборка значений хэш-функций, сравнение и определение результата.

Канонизация текста – это процесс приведения текста к единой форме с помощью удаления из его вспомогательных единиц текста, а также приведения существительных в именительный падеж и единственное число, а иногда и вовсе требуется оставление лишь корневых значений. После этих манипуляций будет получен текст готовый для сравнения.

Шинглы – это упорядоченные множества слов фиксированной длины (длина измеряется в словах), на которые текст «разрезается внахлест». Соответственно, шинглы сохраняют тот же порядок слов, в котором слова следуют в тексте. Разбиение на шинглы представляет собой операцию по разделению текста на последовательности слов длиной от 3 до 10 слов в одном шингле. Главной особенностью является то, что эти последовательности слов идут внахлест, то есть каждый шингл начинается со второго слова предыдущего. Проверка шинглов с количеством менее 3 не имеет смысла, так как похожие словосочетания присутствуют в любом тексте. Количество шинглов можно рассчитать, как количество слов минус длина шинглов и плюс один. Таким образом, чем короче шингл, тем более точным будет результат проверки уникальности.

Алгоритм шинглов представляет собой сравнение случайным образом выбранных значений результатов хэш-функций двух текстов. На этом этапе шинглы вычисляются через хэш-функции, обычно используется 84 хэш-функции (например MD5, SHA1 и прочие) вычисления которых записываются в таблицу. Так весь текст будет представлен в виде двухмерного массива из 84 строк, где каждой строке будет соответствовать хэш-функция. На этапе выборки значений для повышения производительности при сравнении элементов каждого массива необходимо производить случайную выборку результатов хэш-функций для каждой строки. Также можно выбирать значения пожеланию будь то минимальные или максимальные результаты хэш-функций. И на последнем этапе алгоритма сравнивается 84 элемента первого и с соответствующими 84 элементами второго массива, рассчитывается отношение одинаковых значений и вычисляется результат.

УДК 622

Козел А. С.

ПОПУЛЯРНЫЕ СИСТЕМЫ УПРАВЛЕНИЯ КОНТЕНТОМ

БНТУ, г. Минск
Научный руководитель: канд. техн. наук, доцент Дробыш А. А.

В наше время обзавестись собственным Интернет-ресурсом может каждый, однако не все хотят углубляться в основы вёрстки и языки программирования. На помощь данной ситуации приходят системы управления сайтами.

Система управления сайтом представляет собой информационную систему или программу для управления содержимым (контентом). В общем случае системы управления подразделяются на системы управления содержимым масштаба предприятия (Enterprise Content Management System) и системы управления веб-содержимым (Web Content Management