УДК 004

К.А. Лаврентьев,
ассистент кафедры информационных систем и технологий
Хабаровского государственного университета экономики и права,
аспирант ВЦ ДВО РАН
Е.А. Титова

ПРОБЛЕМА ВОЗНИКНОВЕНИЯ КОЛЛИЗИЙ ПРИ ИСПОЛЬЗОВАНИИ ХЕШИРОВАНИЯ ПО СИГНАТУРЕ

Проблема дублирования и нахождения дубликатов в записях базы данных является серьезным тормозящим фактором развития рынка автоматизированных информационных систем. В работе авторами исследован алгоритм нахождения дубликатов в индексированном словаре, проведен анализ быстродействия алгоритма и предложен свой алгоритм.

Ключевые слова: *дублирование, хеширование, алгоритм, сигнатура, коллизия.*

The issue of duplication and finding duplicates in the database records is a serious inhibiting factor in the development of market of automated information systems. In this paper the authors examine the algorithm for finding duplicates in the indexed dictionary, analyze the performance of the algorithm and propose their own algorithm.

Keywords: *duplication, hashing, algorithm, signature, collision.*

Уже давно существует тенденция к созданию программных средств, максимально ориентированная на невнимательность пользователей.

В любом текстовом редакторе существует проверка орфографии и автоматическое исправление, в поисковых системах пользователю в случае опечатки в запросе предлагается исправить ее («возможно вы имели в виду…»). Посредством чего это реализуется? Большинство таких систем основано на алгоритмах нечеткого поиска. Одним из таких алгоритмах является хеширование по сигнатуре.

Хеширование по сигнатуре базируется на достаточно очевидном представлении «структуры» слова в виде битовых разрядов, используемой в качестве хеша (сигнатуры) в хеш-таблице.

Процесс вычисления хеша состоит в следующем: каждому биту хеша сопоставляется группа символов из алфавита. Бит 1 на позиции i в хеше означает, что в исходном слове присутствует символ из i-й группы алфавита. Порядок букв в слове никакого значения не имеет [1].

При использовании данного алгоритма возникают коллизии. Коллизия хеш-функции – это равенство значений хеш-функции на двух различных блоках данных [2]. Если в словаре присутствуют два слова, состоящие из одинаковых букв, то их хеш будет совпадать, ведь в алгоритме хеширования по сигнатуре

порядок букв в слове не имеет значения. По смыслу слова с одинаковыми хешами могут быть абсолютно разными, а значит, теряется суть поиска. В данной статье описана проблема появления коллизий при использования хеширования по сигнатуре.

В описании оригинального алгоритма Л.М. Бойцова используется деление букв алфавита на группы по два, три символа (рисунок 1).

Стоит упомянуть, что в методе, описанном Л.М. Бойцовым, используется 13 битов.

Авторы данной статьи предлагают отличную от оригинального (метод организации хеша Л.М. Бойцова) структуры хеш-таблицы (рисунок 2).

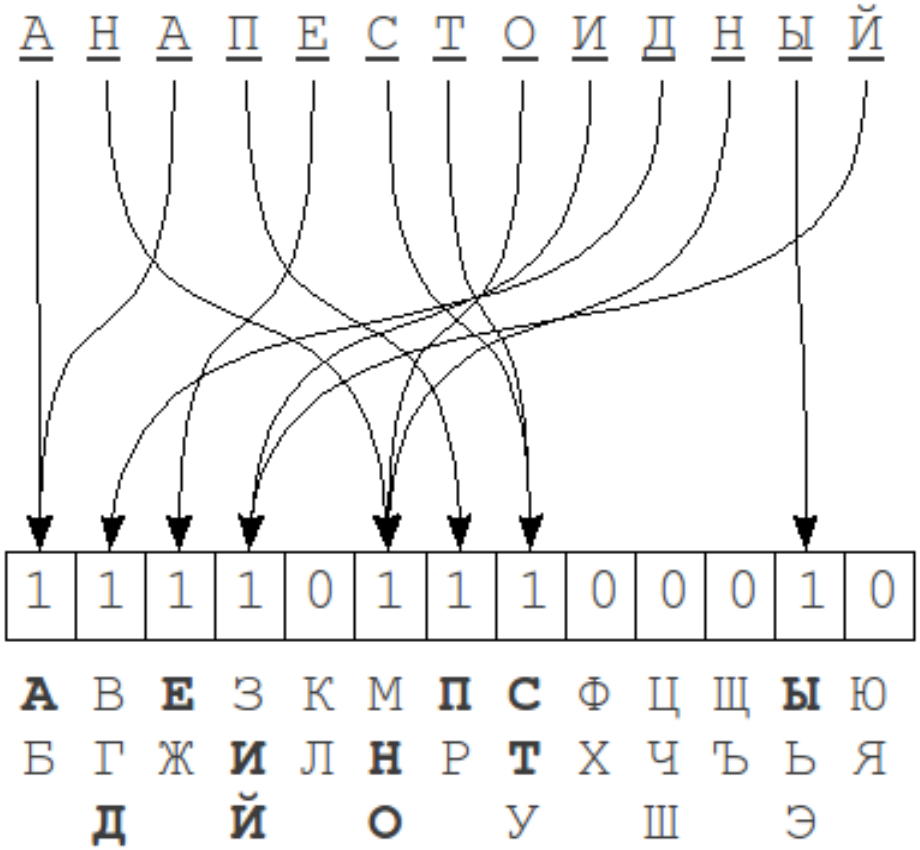

Рисунок 1 – Пример хеширования по сигнатуре

БИТЫ	1	2	3	4	5	6	7	8	9	10	11	12	13
	А	Д	С	Е	Ъ	Ш	Г	Л	Б	Ф	Ж	Ч	Н
	О	Т	З	И	Ь	Щ	К	М	П	Х	Й	Ю	В
				Э	Ы	У	Я			Ц		Р	

Рисунок 2 – Вариант структуры хеш-таблицы

Логика организации данной структуры заключается в принципе парных гласных и согласных букв в русском языке. Действительно, пользователи чаще всего пишут слова так, как произносят. Часто в речи мы заменяем буквы на парные им. 13 битов взято исходя из оригинального алгоритма Л.М. Бойцова. Проведем сравнительный анализ вышеописанных хеш-таблиц. Для сравнения необходимо ввести критерии, по которым будет оцениваться вероятность появления коллизий. Сначала нужно выявить скорость работы алгоритма. Скорость работы будет оцениваться на процессоре Intel® Pentium® CPU 2020M 2.40GHZ. Затем необходимо вывести среднее число коллизий. В данном случае будет использоваться исходный словарь. В качестве словаря в данной статье рассматривается КЛАДР – классификатор адресов России. Расчет процента коллизий производится следующим образом:

1. Находим хеши всех слов в словаре.
2. Для каждого из «хешей» создается список слов-дубликатов, затем они удаляются из исходного словаря (метод «цепочек» разрешения коллизий).
3. Далее находится общее число элементов в списках со-дубликатов и делится на количество слов в словаре.

Результат сравнения показан в таблице.

Таблица – Сравнительный анализ хеш-таблиц

Хеш таблица	**Количество слов в словаре**	**Скорость работы**	**Процент коллизий в словаре**
Хеш-таблица, предложенная Л.М. Бойцовым	10969	0,14 с	83,927 %
Хеш-таблица, предлагаемая авторами	10969	0,13 с	75,795 %

Из данной таблицы можно сделать вывод, что предложенная авторами структура организации хеш-таблицы выигрывает по обоим показателям у оригинальной таблицы Л.М. Бойцова. Также можно говорить о том, что число появления коллизий зависит от структуры хеш-таблицы, и, выбрав подходящий вариант, можно уменьшить количество хешей-дубликатов в словаре.

Список использованных источников

1. https://habrahabr.ru/post/114997/.
2. https://ru.wikipedia.org/wiki/Коллизия_хеш-функции.
3. http://www.intuit.ru/studies/courses/648/504/lecture/11467.