Лабораторная работа №5

Реализация алгоритма обнаружения нечетких дубликатов.

Цель работы: научиться реализовывать на выбранном языке программирования алгоритмы обнаружения нечетких дубликатов на основе алгоритма шинглов.

Краткие теоретические сведения

Проблема обнаружения нечетких дубликатов является одной из наиболее важных и трудных задач анализа данных и поиска информации. Актуальность этой проблемы определяется разнообразием приложений, в которых необходимо учитывать «похожесть», например, текстовых документов — это и улучшение качества индекса и архивов поисковых систем за счет удаления избыточной информации, и объединение новостных сообщений в сюжеты на основе сходства этих сообщений по содержанию, и фильтрация спама (как почтового, так и поискового), и установление нарушений авторских прав при незаконном копировании информации (проблема плагиата или копирайта), и ряд других.
Алгоритм шинглов — алгоритм, разработанный для поиска копий и дубликатов рассматриваемого текста в веб-документе, мощный инструмент, призванный бороться с проявлениями плагиата в интернете.

Реализация алгоритма подразумевает несколько этапов:

- канонизация текстов;
- разбиение текста на шинглы;
- нахождение контрольных сумм;
- поиск одинаковых подпоследовательностей.

Контрольная сумма

В самом общем своем виде контрольная сумма представляет собой некоторое значение, построенное по определенной схеме на основе кодируемого сообщения.

Алгоритм CRC базируется на свойствах деления с остатком двоичных многочленов, то есть многочленов над конечным полем $GF(2^N)$. Значение CRC является по сути остатком от деления многочлена, соответствующего входным данным, на некий фиксированный порождающий многочлен.

Каждой конечной последовательности битов $a_0,a_1,\ldots,a_{N-1}$ взаимно однозначно сопоставляется двоичный полином, последовательность коэффициентов которого представляет собой исходную

последовательность. Например, последовательность битов 1011010 соответствует многочлену:

$$P(x) = 1 * x^6 + 0 * x^5 + 1 * x^4 + 1 * x^3 + 0 * x^2 + 1 * x^1 = x^6 + x^4 + x^3 + x^1$$

Значение контрольной суммы в алгоритме с порождающим многочленом G(x) степени N определяется как битовая последовательность длины N, представляющая многочлен R(x), получившийся в остатке при делении многочлена P(x), представляющего входной поток бит, на многочлен G(x):

$$R(x) = P(x) * x^N \bmod G(x)$$

где

R(x) — многочлен, представляющий значение CRC.

P(x) — многочлен, коэффициенты которого представляют входные данные.

G(x) — порождающий многочлен.

N — степень порождающего многочлена.

Рассмотрим общую схему алгоритма расчета CRC:

1. Выбрать полином P, в результате автоматически становится известна его степень N.
2. Добавить к исходной двоичной последовательности сообщения N нулевых битов. Это добавление делается для гарантированной обработки всех битов исходной последовательности.
3. Выполнив деление дополненной N нулями исходной строки S на полином P по правилам CRC арифметики. Запомнить остаток от деления, который и будет являться CRC.

Задание к лабораторной работе

Написать программу на выбранном языке программирования, реализующую поиск нечетких дубликатов заданных текстов описанным выше алгоритмом. Программа должна запрашивать имена входных файлов и выводить схожие документы и степень их схожести (в процентах).

Контрольные вопросы

1. Что такое нечеткий дубликат?
2. Что такое хеширование?
3. Что такое контрольная сумма?
4. Принципы работы алгоритма шинглов.
5. Принципы работы алгоритма CRC.
6. Что такое CRC-арифметика?
7. Что такое канонизация текстов?

8. Как формировать шинглы?
9. Как выбирать порождающий полином?