АЛГОРИТМ ВОССТАНОВЛЕНИЯ КООРДИНАТ СЛОВ В УСЛОВИЯХ НЕПОЛНОЙ ИЛИ ПОВРЕЖДЁННОЙ ИНФОРМАЦИИ

Белорусский государственный университет информатики и радиоэлектроники
г. Минск, Республика Беларусь

Старостин И. Д.

Гурский Д.И – инженер-программист, ООО "Техартгруп"

Современные системы хранения данных активно используют подход сжатия данных с потерей информации. Одним из примеров может служить формат хранения электронных документов PDF. Спецификация данного формата предусматривает удаление неиспользуемых символов из встраиваемых шрифтов, а также удаление части символьной информации с последующей растеризацией с целью уменьшения объёма файла. Дальнейшее использование текстовой информации становится затруднительным.

Формат PDF хранит дополнительную информацию о тексте и о положении и размере слов в так называемых "глифах" – сущность хранящая текст, координаты и размер текущего слова. После извлечения информации часть слов может оказаться разделённой на несколько отдельных глифов, которые необходимо соединить в целое слово. Для восстановления координат и размера слов предполагается, что:

$$S \geq \lambda \Delta W$$

где S - ширина пробельного символа, W - ширина не-пробельного символа. λ - коэфициент подобия ширины пробельного символа и не пробельного символа для данного шрифта. Примем $\lambda = 0.5$. Данное значение установлено опытным путём, и справедливо для огромного числа различных шрифтов.

На первом шаге алгоритма высчитывается средняя ширина символов.

$$\Delta W = \frac{\sum_{i=1}^{n} \frac{GW_i}{GN_i}}{n}$$

где n - количество глифов, GW_i - ширина i-го глифа, GN_i - количество символов в i-м глифе.

На втором шаге происходит проход по всем существующим глифам с целью поиска тех, которые можно соединить. Глифы, которые необходимо склеить должны удовлетворять следующим условиям:

$$GX_{i+1} - (GX_i + GW_i) \leq \lambda \Delta W$$

$$GY_i \approx GY_{i+1}$$

где GX_i - координата x i-го глифа, GY_i - координата y i-го глифа. λ - коэфициент подобия ширины пробельного символа и не пробельного символа для данного шрифта.

Если глифы удовлетворяют условиям склейки, текст i+1 глифа добавляется к тексту i-го, и происходит пересчёт координат и размеров первого глифа:

$$GX' = GX_i$$

$$GW' = (GX_{i+1} + GW_{i+1}) - GX_i$$

Где GX' и GW' - координата x и ширина нового глифа соответственно.

Проведение анализа реализованного алгоритма показало, что для большинства шрифтов количество успешно восстановленных слов достигает 97%. Доля ошибок при определении слов, которые необходимо восстановить, достигает 1%. Для шрифтов, которые изначально не удовлетворяют условию работы алгоритма, либо показывают плохие результаты, есть возможность регулирования коэффициента λ

Данный алгоритм может быть использован для работы с любым видом текстовых данных. Главным условием является наличие координат x,y, ширины, высоты, и текста конкретного слова. Недостатком можно выделить сильную зависимость от конкретного шрифта, который применяется для данного текста, невозможность использовать на текстовой информации, полученной при использовании разных шрифтов. Однако главным достоинством является отсутствие лексической обработки слова и сопоставления по словарю – это даёт возможность обрабатывать текст на любом языке, а так-же значительно увеличивает производительность по сравнению с методом со словарём.

Список использованных источников:

1. Adobe PDF Reference [Электронный ресурс]. – Электронные данные. – Режим доступа: http://www.adobe.com/content/dam/Adobe/en/devnet/acrobat/pdfs/pdf_reference_1-7.pdf
2. Сметанич, Я. С. Математическая логика, теория алгоритмов и теория множеств / Я. С. Сметанич // Сборник работ. – МИАН СССР, 1973. – 133 с.