УДК 004.896

ДРЕВОВИДНЫЙ АЛГОРИТМ РАЗБИЕНИЯ СЛОВ РУССКОГО ЯЗЫКА НА СЛОГИ НА ОСНОВЕ СТАТИСТИЧЕСКИХ ДАННЫХ

Согоян А.Л.
УО «Брестский государственный технический университет», г. Брест

Огромная часть информации, воспринимаемая человеком, имеет словесный вид. К нему относятся тексты, находящиеся во всемирной паутине, информация, публикуемая в газетах, журналах и других средствах передачи информации. Для того, чтобы тексты выглядели на страницах Интернет-ресурса или газеты в привлекательном и удобном виде, как одно из средств форматирования используют расстановку переносов. В основе расстановки переносов лежит разбиение слов на слоги. В связи с тем, что людям приходится работать с огромными объемами текста, возникает необходимость автоматизировать процесс разбиения слов на слоги – разработать алгоритм, который будет качественно и быстро производить разбиение слова на слоги и, при этом, не будет требовать много памяти.

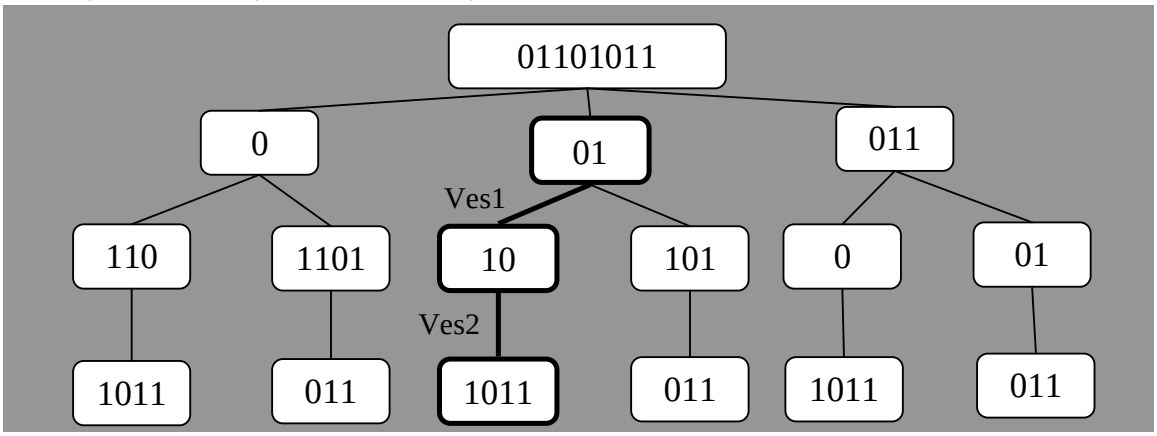
Для сбора весовых коэффициентов цепочек слогов используем имеющийся словарь уникальных русских слов, состоящий из 95725 слов. С помощью существующего онлайн-сервиса quittance.ru производим разбиение всего словаря русских слов на слоги. Данный онлайн-сервис использует в разбиении алгоритм Ляна-Кнута для автоматической расстановки переносов [1].

Полученный словарь русских слов, слоги которых разделены дефисами, преобразуем по правилу замены букв: гласные заменяем нулем, согласные единицей. Формируем массив вхождений соседних слогов в слова словаря.

В результате статистического анализа словаря русских слов были получены, для цепочки из 2 слогов – 289 вариантов связей, для цепочки из 3 слогов – 1092. Для цепочки из двух слогов наибольшую частоту имеет «10-10», а для цепочки из трех слогов «10-10-10».

На вход функции подается слово из русского языка. Буквы в слове заменяются цифрами 1 или 0. Для согласной буквы – 1, для гласной – 0. В результате имеется бинарное слово, которое представляется в виде дерева слогов. На рисунке изображено дерево для разбиения на слоги слова «алгоритм». В бинарном представлении данное слово будет выглядеть как «01101011». Количество нулей в бинарном слове соответствует количеству уровней в дереве, не считая самого верхнего.

Получив дерево всех возможных разбиений слова, начинаем его обход по каждой ветви. Имея статистическую матрицу связей слогов, считаем вес всего разбиения. Производим суммирование весов связей слогов на соседних уровнях. Разбиение с максимальным весом имеет наибольшую вероятность[2]. Исходя из этого, за правильное разбиение принимается разбиение бинарного слова на слоги с наибольшим весом.



Количество \ Длина цепочки	**2**	**3**
Слов	95 725	95 725
Правильных разбиений	53 623	64 607
Неправильных разбиений	42 102	31 118
Эффективность	56,02%	67,49%

Используя исходный словарь русских слов, на основе которого формировался массив коэффициентов связей цепочек слогов, произведем разбиение слов словаря и сравним с эталонным разбиением. В таблице представлены данные по количеству слов, участвующих в разбиении на слоги, количеству правильных результатов, а также эффективности алгоритма, использующего цепочки слогов длиной 2 и 3.

Среди слов, которые не были правильно разбиты программой на основе описываемого алгоритма, преобладают имеющие два или более корня в своем составе. С увеличением длины цепочки эффективность древовидного алгоритма разбиения слов русского языка растет. В связи с тем, что алгоритм не дает точного результата для всех слов, его целесообразно использовать в комбинации с другими методами. Так, например, использовать одновременно разбиения слов при вычислении веса разбиения с цепочками длиной два, три и более. А как результат – выбрать некоторую оптимальную комбинацию.

Литература

1. Donald E. Knuth. Digital typography. CSLI Lecture Notes, no. 78. Stanford, 1999.
2. Яглом, А.М. Вероятность и информация / А. М. Яглом, И. М. Яглом – М.: Наука, 1973.

УДК 004.514.62

ИСПОЛЬЗОВАНИЕ ПЕРИФЕРИЧЕСКОГО ЗРЕНИЯ ПРИ НАВИГАЦИИ МЕЖДУ СТРАНИЦАМИ В ИНТЕРНЕТ-БРАУЗЕРЕ

Тавониус К.А.
УО «Брестский государственный технический университет», г. Брест

В [1] приведено сравнение с лабиринтом ориентирования в современном программном интерфейсе, когда пользователь не имеет возможности видеть одновременно, хотя бы схематично, изображение всего рабочего пространства. В основном такой подход вызван ограниченностью аппаратных ресурсов персонального компьютера, не позволяющих задействовать большие площади для вывода информации.

В последнее время все большее внимание уделяется попыткам использовать уменьшенный масштаб изображений, не находящихся в фокусе работы пользователя, для увеличения наглядности и интуитивности интерфейса [2]. Не в последнюю очередь оживление в данной области связано с ростом разрешающей способности дисплеев, делающей более информативной технологию применения уменьшенных изображений объектов для предварительного просмотра (previews или thumbnails).

В последнее время масштабные преобразования начинают применяться в средствах Интернет-навигации. Однако их использование на сегодняшний день ограничено использованием сильно уменьшенных изображений веб-страниц в качестве ярлыков. В данной работе представлено дальнейшее развитие модели использования масштабирования веб-страниц, основанное на использовании аналогии периферического зрения.