УДК 681.3
Р.Д. Миронюк
Тернопільський національний економічний університет, Україна

АЛГОРИТМ ВИДІЛЕННЯ КЛЮЧОВИХ СЛІВ КОНТЕНТУ WEB-САЙТІВ НА ОСНОВІ SEO ОПТИМІЗАЦІЇ

R.D. Myronyuk

ALGORITHM OF SELECTION OF KEYWORDS OF THE CONTENT OF WEB SITES BASED ON SEO OPTIMIZATION

У багатьох випадках, пошук необхідної інформації в на веб сторінках здійснюється за ключовими словами, які визначаються автором і відображають основні положення сторінок. Завдання ключових слів визначається на етапі внесення інформації про веб сторінку і є досить відповідальним моментом. Коли інформація про сайт включається у відповідь на запит при наявності певних ключових слів, пов'язаних з сайтом і заданих розробником сайту. Як видно проблема полягає в необхідності попереднього завдання набору ключових слів і відсутність деяких з них виключає веб сторінку зі списку пошуку

У даній роботі пропонується підхід для пошуку ключових слів контенту для кожної веб сторінки. Підхід будується на основі підходу аналогічного побудови дискретного вейвлет перетворення і згладжування за методом ковзного відомості

Уявімо текст з файлу в вигляді ряду {x (n)} довжиною L (n = 0,1, ..., L-1). Текст доцільно привести до нижнього регістра і видалити з нього прогалини, знаки пунктуації та інші незначні символи. Значеннями x (n) можна вважати, наприклад ASCII-коди символів (хоча можна використовувати і власну систему кодування, наприклад, числа з діапазону [-1, + 1]). На представленому малюнку 1 на верхньому графіку зображений саме такий текстовий ряд {x (n)}.

Будемо вважати, що ключове слово, що має довжину N, присутній в тексті і розташоване, починаючи з позиції n0.

Вейвлет перетворення являє собою інтегральне перетворення вихідного сигналу за допомогою вейвлет функцій. Воно дещо схоже на перетворення Фур'є, але при цьому дозволяє локалізувати частотні зміни сигналу в часі.

Накладаючи вихідний сигнал на масштабований вейвлет і проводячи інтегрування по всій тимчасової області отримують нове двомірне подання про вихідному сигналі. Нове уявлення в ряді випадків дозволяє виявити певні закономірності в сигналі, зробити його ущільнення і фільтрацію. Дискретне вейвлет перетворення задає нове уявлення сигналу, що складається з усереднений і деталізацій.

У даній роботі пропонується з вихідного сигналу отримати «карту усереднений». Для її отримання будемо розглядати прямокутний «вейвлет» пременной довжини N, який будемо переміщати вздовж сигналу. Як видно з малюнка при досягненні відстані n0 відбудеться накладення «вейвлета» на ключове слово і, отже, воно буде виявлено. Таким чином, необхідно вибрати діапазон ширини прямокутника «вейвлета» і спосіб обчислення «вейвлет перетворення». Як спосіб обчислення

«Вейвлет перетворення» доцільно обчислювати середнє значення для діапазону попавшого в середину «вейвлета». Даний процес аналогічний обчисленню за методом змінного середнього. Слід почати з трибуквених усереднений, далі отримати чотирьох-буквені, потім п'яти-буквені і т.д. до восьми-буквених усереднений. Знаходити довші усереднення мабудт недоцільно, тому що довжина більшості слів зосереджена в цьому

діапазоні. Таким чином, буде отримана деяка карта (матриця) складається з 6 рядків і L-2 стовпців, причому стобци L-3, L-4, ..., L-7 в нижніх рядках дорівнюють нулю.

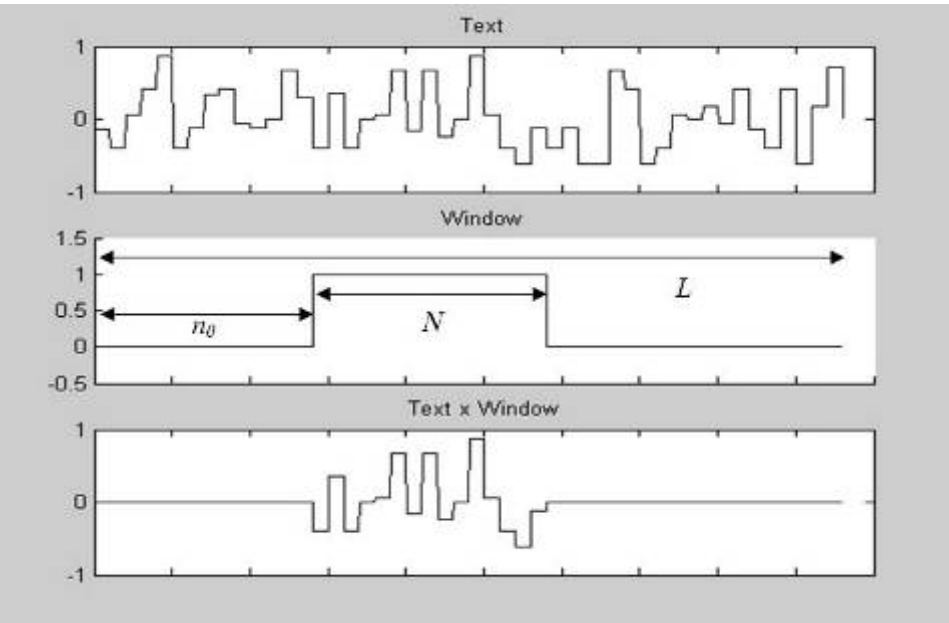

Рисунок 1. Подання тексту у вигляді сигналу {x (n)}

У ряді випадків можуть бути помилки. Наприклад, слова «МИР» і «РИМ» матимуть однакові середні значення, але при реальному пошуку зазвичай задаються досить довгі слова (словосполучення) і ймовірність збігів стає невеликий. При цьому слід зазначити, що краще видати надлишкову інформацію, ніж пропустити будь-якої документ.

Аналіз сучасних пошукових систем, що працюють з текстом, показує, що вони будуються на індексування тексту, тобто побудові спеціального словника, що складається з усіх слів веб сторінки. Сама по собі процедура побудови такого словника є дуже громіздкою і вимагає великої кількості ресурсів.

Пропонований метод працює з картою усереднення тексту, в якій міститься в новому якісному вигляді вміст текстового файлу

Література

1. Нейроподібні методи, алгоритми та структури обробки сигналів і зображень у реальному часі: монографія / Ю.М. Рашкевич, Р.О. Ткаченко, І.Г Цмоць, Д.Д. Пелешко. – Львів: Видавництво Львівської політехніки, 2014. -256 с.

2. Проблемно–ориентированные высокопроизводительные вычислительные системы: В.Ф. Гузик, В.Е. Золотовский: Учебное пособие. Таганрог:Изд-во ТРТУ, 1998. 236 с.

3. Уоссермен Ф. Нейрокомпьютерная техника. – М.: Мир,1992. – 259с.