Wyszukiwanie informacji w sieci, Lista 2 **10.03.2007**

(zadania na ćw. w tygodniu 12.03–18.03.2007)

1. [1] Rozważmy następujące uogólnienie odległości edycyjnej słów nad alfabetem Σ.

   - dla każdej litery $a \in \Sigma$, wstawienie oraz usunięcie a ze słowa mają koszty $ins(a)$ oraz $del(a)$;
   - dla każdej pary liter $a, b \in \Sigma$ i $a \neq b$, koszt zamiany litery a na b wynosi $rpl(a, b)$;

   Podaj algorytm, który w czasie $O(n \cdot m)$ wyznacza uogólnioną odległość edycyjną słów o długościach n i m.

2. [1] Aby ograniczyć liczbę rozważanych termów w wyszukiwaniu przybliżonym wg odległości edycyjnej, wyznaczanie odległości edycyjnej stosuje się jedynie dla termów o małej odległości Jaccarda od zapytania.

   Rozważmy tylko słowa o tej samej długości n i współczynnik Jaccarda dla bigramów. Spróbuj oszacować największą i najmniejszą możliwą odległość edycyjną słów w i v, dla których współczynnik Jaccarda wynosi c, dla $c = 0$ oraz dla $c = 1$.

3. [1] Opracuj "polski odpowiednik" algorytmu Soundex: podziel litery, grupy liter odpowiadające głoskom na rozłączne "klasy abstrakcji". Podaj argumenty za/przeciw zachowaniu ogólnego schematu algorytmu Soundex. Omów najbardziej kłopotliwe przypadki.

4. [1] Jako alternatywa dla indeksu odwróconego, w literaturze pojawiają się tzw. pliki sygnaturowe (signature files). Zapoznaj się z tą strukturą, przedstaw jej wady/zalety w porównaniu z indeksami odwróconymi.

   Omów też na czym polega modyfikacja signature files nazywana bitsliced signature files.

5. [0.5] Na wykładzie omówione zostało indeksowanie rozproszone ze scalaniem bloków, wyglądające dość podobnie do klasycznego algorytmu sortowania przez scalanie. Dlaczego nie stosuje się zatem standardowej wersji sortowania przez scalanie?

6. [1] Przyjmijmy następujące parametry sprzętu komputerowego:

oznaczenie	parametr	wartość
s	dostęp do dysku (zmiana pozycji głowicy)	10^{-2} s
b	czas transferu danych z/na dysk (na 1byte)	10^{-6} s
p	inne operacje procesora (np. porównanie/zamiana słów)	10^{-7} s.

   Zindeksować chcemy kolekcję dokumentów o następującej wielkości:

oznaczenie	parametr	wartość
N	liczba dokumentów	10^9
L	liczba termów w dokumencie	10^3
M	liczba różnych termów	$44 \cdot 10^6$

   Podaj czas następujących etapów indeksowania:

	krok	czas
1	czytanie danych z dysku	
2	quicksort dla bloków mieszczących się w RAM	
3	zapis posortowanych bloków na dysk	
4	czas operacji dyskowych przy scalaniu	
5	razem	

Wybierz rozmiar sortowanego bloku adekwatny do dzisiejszych możliwości sprzętowych.

7. [0.5] Aby uniknąć dodatkowego przebiegu przez indeksowane pliki, słownik w procesie indeksowanie powinien być tworzony na bieżąco, w możliwie małej pamięci (aby zmieścił się w RAM) z możliwością wyszukiwania. Zaproponuj jakieś własne rozwiązanie.

8. [1] Dla problemu jak w poprzednim zadaniu omówiono różne rozwiązania i zaproponowano nowe w pracy:

   "Burst tries: a fast, efficient data structure for string keys", S. Heinz, J. Zobel, and H.E. Williams, ACM Transactions on Information Systems, 20(2):192-223, 2002.

   Omów przedstawione tam rozwiązania, w tym burst tries.

9. [1] Rozważmy indeksowanie zbioru n dokumentów, każdy dokument składa się z m termów. Przyjmijmy też, że liczba różnych termów (czyli rozmiar słownika) wynosi t.

   (a) Wyznacz wartość stałej c z prawa Zipfa dla tej kolekcji dokumentów i liczbę wszystkich adresów dokumentów na listach indeksu odwróconego.

   (b) Stosując prawo Zipfa, oszacuj wymagania pamięciowe dla indeksu odwróconego w postaci list numerów dokumentów, w których występują termy (numery dokumentów kodowane przy pomocy kodów o stałej długości).

   (c) Przyjmijmy teraz regularne "rozproszenie" dokumentów na liście każdego termu: odległości między sąsiednimi numerami dokumentów termu występującego w (mniej więcej) co i-tym dokumencie wynoszą n/i. Wyznacz wymagania pamięciowe przy tym założeniu, gdy na liście sąsiadów termu kodowane są tylko: numer pierwszego dokumentu, i różnice między numerami sąsiednich dokumentów. Zastosuj kodowanie γ!

   Podaj wielkości liczbowe wyników dla powyższych wariantów, gdy $n = 10^6$, $m = 10^3$ a $t = 10^5$.