

Raport

18 maja 2009

Spis treści

1	Opis formatu słownika morfologicznego	2
1.1	Format pliku	2
1.2	Format w programie	2
2	Indeks odwrócony	2
2.1	Format w programie	2
2.2	Format listy postingowej	2
2.3	Tworzenie indeksu	3
3	Stemming	3
4	Wyniki	3
4.1	Wielkość indeksu	3
4.2	Czasy wyszukiwania	3
4.3	Pomiar czasu i pamięci dla indeksu skompresowanego ze stemmingiem	3
4.4	Pomiar czasu i pamięci dla indeksu nieskompresowanego ze stemmingiem . .	4
4.5	Pomiar czasu i pamięci dla indeksu skompresowanego bez stemmingu	4

1 Opis formatu słownika morfologicznego

Słownik morfologiczny ma format zmieniony w stosunku do wersji oryginalnej.

1.1 Format pliku

Plik binarny słownika morfologicznego ma następujący format:

Długość listy form bazowych ze słownika jako uint.

Lista form bazowych jako ciąg stringów o długości zapisanej wcześniej. Ilość par (słowo, lista numerów form bazowych) ze słownika. Dla każdej pary takiej jak powyżej słowo jako string, następnie długość listy numerów form bazowych, zaś dalej lista tych numerów.

1.2 Format w programie

Po wczytaniu z pliku (w trakcie działania programu) słownik morfologiczny jest przechowywany jako seria tablic.

- Posortowana tablica słów, dla których można odnaleźć formy bazowe (wyszukiwanie binarne).
- Tablica tablic numerów form bazowych. Te numery odpowiadają indeksom w tablicy form bazowych. Natomiast indeksy są takie jak dla tablicy słów powyżej.
- Tablica form bazowych.

Wczytywanie słownika morfologicznego trwa: 00:00:04.3992.

Użyta pamięć: 229119604 bajtów.

2 Indeks odwrócony

2.1 Format w programie

W indeksie przechowujemy:

- Posortowaną tablicę zindeksowanych tokenów (słownik).
- Tablicę list postingowych o indeksach odpowiadających indeksom słów.
- Tablicę pozycji początków kolejnych artykułów w pliku.

Słowa wyszukujemy binarnie w tablicy tokenów.

2.2 Format listy postingowej

W liście postingowej przechowujemy:

- Tablicę pozycji dokumentów (pozycja identyfikuje dokument)
- Tablicę tablic pozycji w dokumencie.

W wersji skompresowanej lista postingowa składa się pojedynczej wartości int - oznaczającej liczbę dokumentów, w których występuje dany term oraz z tablicy bajtów. W tablicy zapisujemy przy pomocy kompresji gamma kolejno: identyfikator dokumentu *id*, liczbę wystąpień danego termu w dokumencie *id*, kolejne pozycje termu w tym dokumencie.

2.3 Tworzenie indeksu

Indeks odwrócony tworzymy w dwóch przebiegach.

W pierwszym przebiegu tworzymy posortowany słownik mapujący słowa na listy postingowe. Gdy skończy się miejsce zapisujemy go do pliku tymczasowego i tworzymy nowy słownik. Po przetworzeniu całego pliku źródłowego łączymy powstałe pliki tymczasowe do jednego wynikowego.

W drugim przebiegu wyliczamy pozycje dokumentów w pliku z Wikipedią i zapisujemy do pliku.

3 Stemming

Nasz program wykorzystuje stemming przy tworzeniu indeksu oraz wyszukiwaniu.

Najpierw obcinane są z końca słowa rozpoznane końcówki które można usuwać. Następnie z końca słowa usuwamy samogłoski. W przypadku gdy nie ma więcej końcówek do usunięcia lub słowo robi się zbyt krótkie stemming jest zakończony.

Stemming pozwala na zmniejszenie słownika o ok. 16%.

4 Wyniki

Przeprowadziliśmy testy naszego programu na wybranych opcjach tworzenia indeksu oraz wybranych zapytaniach.

4.1 Wielkość indeksu

- **Wielkość słownika** ze stemmingiem 2073396, bez wykonania stemmingu: 2473993.
- **Zindeksowanych dokumentów** 811205

4.2 Czasy wyszukiwania

Poniżej czasy wyszukiwania dla zbiorów zapytań.

- **Zapytania AND:** całkowity czas przetwarzania: 00:00:47.8452000
- **Zapytania OR:** całkowity czas przetwarzania: 00:02:14.0508000
- **Zapytania frazowe:** całkowity czas przetwarzania: 00:00:55.5048000

4.3 Pomiar czasu i pamięci dla indeksu skompresowanego ze stemmingiem

- Czas tworzenia indeksu: 01:11:40.1556
- Czas wczytywania: 00:00:16.2864000
- Zajmowana pamięć: 437 MB
- Skompresowane listy postingowe: tablice bajtów: 298 MB (68,24%) reszta: 47 MB (10,85%)

- Słownik: 69 MB (15,76%)
- Tablice String[] 7,9 MB(1,81%)
- Tablice PositionalPostingList[] 7,9 MB(1,81%)
- Tablice Int64[] 6,2 MB (1,42%)

4.4 Pomiar czasu i pamięci dla indeksu nieskompresowanego ze stemmowaniem

- Czas tworzenia indeksu: 01:08:50.2248
- Czas wczytywania: 00:01:08.5152000
- Zajmowana pamięć: 1.5 GB
- Listy postingowe: 1.4 GB (94.02%) w tym:
 - indeksy dokumentów 253 MB
 - listy pozycji $886 + 261 = 1147$ MB
- Słownik: 69 MB
- Tablice PositionalPostingList[] 32 MB (2,08%)
- Tablice Int64[] 6,2 MB

4.5 Pomiar czasu i pamięci dla indeksu skompresowanego bez stemmowania

- Czas tworzenia indeksu: 00:58:41.6688
- Czas wczytywania: 0:00:18.8136000
- Zajmowana pamięć: 490 MB
- Skompresowane listy postingowe: tablice bajtów: 323 MB (65,91%) reszta: 57 MB (11,57%)
- Słownik: 84 MB (17,21%)
- Tablice String[] 9,4 MB(1,93%)
- Tablice PositionalPostingList[] 9,4 MB(1,93%)
- Tablice Int64[] 6,2 MB (1,26%)