

PAMSI - Sprawozdanie 5

Filip Guzy 218672

23 kwietnia 2016

Tablica asocjacyjna

Tablica asocjacyjna jest strukturą danych w której odwołanie do elementu jest możliwe za pomocą przypisanego mu klucza. Istnieje kilka różnych sposobów implementacji tej struktury danych. Jedną z najpopularniejszych jest struktura oparta na tablicy haszującej, czyli wykorzystującej specjalną metodę kodującą klucze (najczęściej na integerzy). W celu doświadczalnym stworzono tablicę asocjacyjną opartą na tablicy haszującej, w której metoda odpowiedzialna za haszowanie wykorzystuje kombinację przesunień bitowych w celu zamiany trzech pierwszych liter klucza na dodatnie liczby całkowite. Wykorzystano również mechanizm tzw. 'rehashowania': po każdym wypełnieniu przynajmniej jednym elementem wszystkich dostępnych w tablicy list następuje podwojenie rozmiaru tablicy i ponowne rozdzielenie elementów przy wykorzystaniu metody haszującej. Umożliwi to (w pewnym stopniu) zmniejszenie czasu odczytu pewnych elementów kosztem dłuższego ich zapisu.

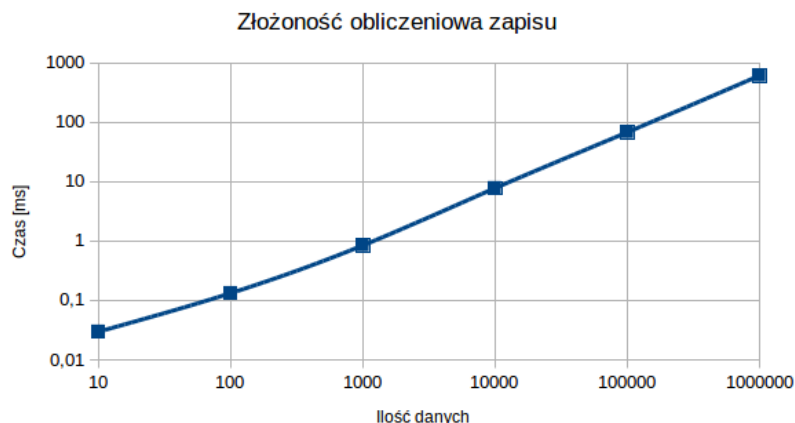
Pomiary zapisu i odczytu elementów

W celu sprawdzenia złożoności obliczeniowej operacji zapisu elementów do tablicy wykorzystano książkę telefoniczną zawierającą milion nazwisk i przyporządkowanych im numerów telefonów. Numery telefonów były kolejno wczytywane i zapisywane do tablicy według klucza - nazwisk. Zgodnie z teorią zapis jednego elementu powinien być wykonywany w czasie stałym, czyli jego złożoność powinna wynosić $O(1)$, natomiast złożoność zapisu pewnej ilości elementów - $O(n)$. Poniżej przedstawiono pomiary czasu (w ms) wykonane dla zapisu 10, 100, 1000, 10000, 100000 oraz 1000000, czyli największej możliwej wartości biorąc pod uwagę pojemność książki telefonicznej (wykorzystanej do wczytywania danych). Przyjęto tablicę list o początkowym rozmiarze równym 10, która ulegała powiększaniu i rehashowaniu z każdym przepełnieniem.

Tabela 1: Zapis elementów do tablicy asocjacyjnej

Ilość	10	100	1000	10000	100000	1000000
Czas [ms]	0,033	0,147	0,851	8,674	62,276	608,983
	0,028	0,128	0,831	8,109	74,073	629,731
	0,028	0,127	0,837	8,05	69,662	606,956
	0,029	0,128	0,84	6,32	64,193	604,452
	0,0289999	0,13	0,884	7,917	65,542	612,631
	0,029	0,13	0,839	8,06	71,568	604,171
	0,03	0,134	0,844	8,223	66,415	602,449
	0,0289999	0,128	0,841	6,397	68,39	604,389
	0,0320001	0,145	0,84	8,067	65,548	606,843
	0,03	0,127	0,826	7,97	71,676	621,482
Średnia	0,02969999	0,1324	0,8433	7,7787	67,9343	610,2087

Pomiary przedstawiono także na poniższym wykresie:



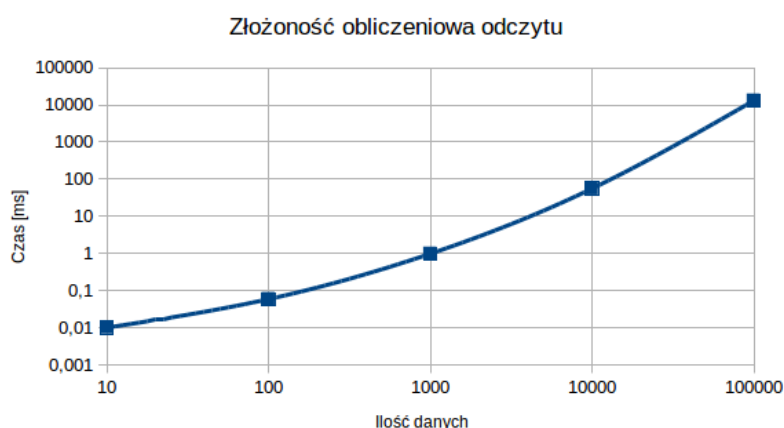
Jak można zauważyć, zgodnie z założeniem wraz ze wzrostem danych dziesięciokrotnie, tyle samo wzrasta czas wykonania zapisu, zatem złożoność obliczeniowa dla zapisu serii danych zawiera się w $O(n)$, natomiast w przypadku zapisu pojedynczego elementu - $O(1)$. Operacja rehashowania ma niewielki wpływ na zmianę złożoności całego zapisu, lecz przyczyna tego zostanie dokładniej wyjaśniona w opisie odczytu.

W następnej kolejności zmierzono czas odczytu danych. Zgodnie z teorią oraz optymistycznym założeniem odczyt jednego elementu powinien mieć złożoność $O(1)$, natomiast w przypadku średnim $O(n)$. Przeszukiwanie książki telefonicznej w celu znalezienia pewnej ilości nazwisk powinno więc dokonywać się w czasie kwadratowym. W poniższej tabeli przedstawiono czasy (w ms) wyszukiwania tych samych ilości danych co w przypadku zapisu.

Tabela 2: Odczyt elementów z tablicy asocjacyjnej

Ilość	10	100	1000	10000	100000
Czas [ms]	0,011	0,059	1,011	57,908	12403,7
	0,00900006	0,0589999	0,967	55,553	13083,1
	0,00900006	0,0580001	0,969	53,873	13426,9
	0,00999999	0,058	0,965	57,85	12407,8
	0,00999999	0,059	0,989	51,121	13166,1
	0,00999999	0,059	0,974	54,944	13757,4
	0,012	0,0580001	1,004	59,175	13594,5
	0,00999999	0,0589999	0,97	58,639	13560,7
	0,00900006	0,059	0,97	51,766	12550
	0,00999999	0,059	0,969	58,176	12493,1
Średnia	0,010000013	0,0587	0,9788	55,9005	13044,33

Pomiary przedstawiono także na poniższym wykresie:



Obserwując zmierzone dane można wywnioskować, że czas odczytu pewnej ilości danych jest rzędu $C * n^2$, gdzie C jest pewną wartością zależącą od doboru metody haszującej i różnorodności kluczy. W przypadku przepełnienia tablicy następuje jej powiększenie i ponowny rozdział elementów, jednak przy wielu powtórzeniach kluczy (w wykorzystanej książce telefonicznej nazwiska powtarzają się wielokrotnie) tablica ulega powiększeniu jedynie raz lub dwa razy, co implikuje dodawanie kolejnych elementów do już istniejących list, jednocześnie zwiększając wartość C i złożoność obliczeniową, co jest widoczne w przedstawionej tabeli i na wykresie. Dla większej różnorodności kluczy C powinna mieć wartość zbliżoną do 1.

Wnioski

Tablice asocjacyjne bazujące na tablicach z hashowaniem są idealną strukturą dla sparowanych danych. Zapis lub odczyt takich tablic ma złożoność obliczeniową zależną od intencji programisty, można bowiem usprawnić zapis kosztem odczytu (przykładem może być sortowanie list podczas zapisu, co zwiększa jego czas zmniejszając tym samym czas odczytu) lub odczyt kosztem zapisu (zapis dokonywany w czasie stałym, jednak utrudniony odczyt). W przeprowadzonym doświadczeniu został wykorzystany drugi sposób. Otrzymane złożoności obliczeniowe nie odzwierciedlają idealnie założeń teoretycznych, jednak są im zbliżone.