

Badanie różnych implementacji tablic asocjacyjnej pod kątem dostępu do jej elementów

Szymon Leśniak

28 maja 2014

1 Wstęp

Tablica asocjacyjna jest strukturą danych, w którą zbiera pary danych, niekoniecznie tych samych typów: klucza, który zwykle jest wartością stałą, oraz wartości, dowolnie modyfikowalnej. Tablice asocjacyjne implementowane są tak, by minimalizować czas dostępu (modyfikacji lub pobrania) wartości. Jest to operacja na tyle popularna, że C++ STL definiuje ją za pomocą operatora `[]` (szablon `std::map`).

2 Opis badania i szczegóły implementacji

Do badania przeznaczonych było 5 zestawów niepowtarzających się napisów w różnej ilości: 100, 1000, 10000, 50000 i 100000 sztuk. Każdym z zestawów wypełniano tablicę asocjacyjną. Mierzony był czas wpisania wartości do losowo wybranego elementu tablicy. Operację powtarzano wielokrotnie w celu zniwelowania błędów przypadkowych.

Badanymi implementacjami były: prosta tablica dynamiczna, drzewo binarne (bez mechanizmów równoważenia długości gałęzi) oraz tablicę haszującą. Funkcja haszująca była oparta na operacji modulo. Ostatnia implementacja była badana w trzech wariantach, odpowiadających różnemu wypełnieniu tablicy: 10%, 50% i 90%. Konieczne było zatem odejście od spotykanego zalecenia, by rozmiar wyrażał się liczbą pierwszą.

3 Wyniki badania

Wyniki są przedstawione w poniższej tabeli oraz na wykresie (skala log-log):

4 Wnioski

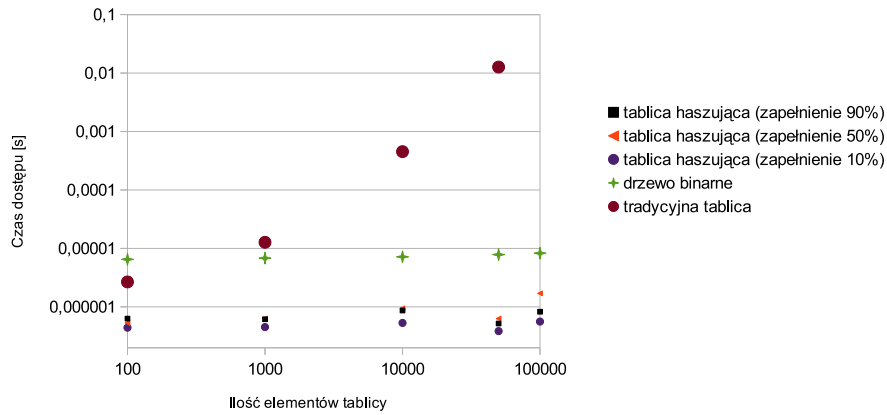
Zgodnie z oczekiwaniami zwykła tablica dała najgorszy czas dostępu do danych. Dostęp ma złożoność $O(n)$ i jako taki jest nieakceptowalny dla większych zbiorów danych.

Tablica 1: Czasy dostępu do elementów tablicy asocjacyjnej (w sekundach)

Ilość próbek	tradycyjna tablica	drzewo binarne	tablica haszująca (zapełnienie 10%)	tablica haszująca (zapełnienie 50%)	tablica haszująca (zapełnienie 90%)
100	$2,67 \cdot 10^{-6}$	$6,50 \cdot 10^{-6}$	$4,41 \cdot 10^{-7}$	$5,31 \cdot 10^{-7}$	$6,30 \cdot 10^{-7}$
1000	$1,27 \cdot 10^{-5}$	$6,82 \cdot 10^{-6}$	$4,52 \cdot 10^{-7}$	$6,34 \cdot 10^{-7}$	$6,14 \cdot 10^{-7}$
10000	0,000452823	$7,17 \cdot 10^{-6}$	$7,17 \cdot 10^{-7}$	$9,45 \cdot 10^{-7}$	$8,67 \cdot 10^{-7}$
50000	0,0126578	$7,84 \cdot 10^{-6}$	$7,84 \cdot 10^{-7}$	$6,32 \cdot 10^{-7}$	$5,20 \cdot 10^{-7}$
100000		$8,29 \cdot 10^{-6}$	$8,29 \cdot 10^{-7}$	$1,70 \cdot 10^{-6}$	$8,26 \cdot 10^{-7}$

Średni czas dostępu do elementów tablicy asocjacyjnej

w zależności od jej implementacji



Potwierdzona została przewidziana złożoność dostępu $O(1)$ dla tablicy haszującej. Duży rozmiar tablicy przyspiesza nieznacznie średni czas dostępu do danych, co wynika z mniejszego prawdopodobieństwa wystąpienia kolizji.

Jedynie dla drzewa binarnego zgodność z założeniami teoretycznymi (tj. złożoność $O(\log n)$) jest widoczna w niewielkim stopniu.