Д. И. Синица

ANALYSIS OF HASHING WITH OPEN ADDRESSING. ADVANTAGE OF DOUBLE HASHING

The article deals with the assesment of hashing with open addressing under the assumption of uniform hashing. Much attention is given to the sufficient condition of double hash function for the entire tableto be searched.

Introduction

Double hashing offers one of the best methods available for open addressing. It uses a function

$$h(k,i) = (h_1(k) + ih_2(k)) \bmod m,$$

where m is the hash table size, k is the key of the element, $i \in \{0,1,...,m-1\}$ is the probe number, both $h_1(k)$ and $h_2(k)$ are auxiliary hash functions.

I. Sufficient condition of full hash table coverage

Theorem 1 If the value of $h_2(k)$ in the double hash function is relatively prime to the hash-table size m then the probes cover the entire table.

Proof Choose without loss of generality a random key k_r from the universe of keys U.

$$h(k_r,i) = (h_1(k_r) + ih_2(k_r)) \bmod m,$$

$$h_1(k_r) = r_1, h_2(k_r) = r_2,$$

$$h(i) = (r_1 + ir_2) \bmod m,$$

where m and r_2 are relatively prime.

The resulting function examines the entire hash table only in case if it accepts all the values in the set $\{0,1,...,m-1\}$ once, while $i = 0,1,...,m-1$.

Let's assume there are i_1 and i_2 such that

$$(r_1 + i_1r_2) \bmod m = (r_1 + i_2r_2) \bmod m = s.$$

Then rewrite this equals in another form.

$$r_1 + i_1r_2 = q_1m + s,$$

$$r_1 + i_2r_2 = q_2m + s.$$

where q is the integer part of the ratio of $r_1 + ir_2$ and m. The difference of these expressions is

$$r_2(i_1 - i_2) = m(q_1 - q_2).$$

As the right side is exactly divisible by m, therfore the left one as well, but $i_1 - i_2 < m$ and q doesn't divide by m without remainder (relatively prime). So the original proposition must be true.

II. Designing double hash function covering the entire table

A convenient way to ensure the sufficient condition of full table coverage is to let m be prime and to design $h_2(k)$ so that it always returns a positive integer less than m.

$$h_1(k) = k \bmod m,$$

$$h_2(k) = 1 + (k \bmod (m-1)).$$

In this case double hashing improves over other approaches by usage of $\theta(m^2)$ probe sequences, since each possible $h_1(k)$, $h_2(k)$ pair yields a distinct probe sequence.

III. Analysis of open-address hashing

The expected number of probes for an open-address hash table with load factor $\alpha = n/m$ under the assumption of uniform hashing is
Theorem 2.1 $\frac{1}{1-\alpha}$ in an unsuccessful search.
Theorem 2.2 $\frac{1}{1-\alpha}$ for inserting an element.
Theorem 2.3 $\frac{1}{\alpha}\ln(\frac{1}{1-\alpha})$ in an successful search.

4* Load percent	2*Operations		
	Succesful search	2*Insertion	Unsuccesful search
2*50%	1.365 (1.386)	2 (2)	2.034 (2)
2*70%	1.753 (1.72)	3.323 (3.333)	3.239 (3.333)
2*90%	2.525 (2.558)	9.838 (10)	10.262 (10)

The table above shows experimental and calculated values for those operations.

IV. Conclusion

For hash function described in the 2nd paragraph the performance appears to be very close to the "ideal" scheme of uniform hashing.

1. Introduction to algorithms / Thomas H. C. [et all.] // MIT – 2009 – P. 253–286.

Синица Дмитрий Игоревич, студент 2-го курса, ФИТиУ, БГУИР man.upon.problems@gmail.com.

Научный руководитель: Навроцкий Анатолий Александрович, зав. кафедрой информационных технологий автоматизированных систем, кандидат физ.-мат. наук, доцент, navrotsky@bsuir.by.