Reuse Distance Distribution in Random Access

Xiaoming Gu
Department of Computer Science
University of Rochester
xiaoming@cs.rochester.edu

URCS Technical Report #930
January 29, 2008

Abstract

Reuse distance is a basic metric for program locality. The distribution of reuse distances, called the reuse signature, shows the average locality or the amount of actively used data. Random access is often assumed in analytical models about program behavior. An interesting question is whether the reuse behavior of random data access has a closed-form answer. In this paper we prove that the length of reuse distances of random access is uniformly distributed from 0 to $n - 1$ when $n$ is the size of data. We also test random traces of different lengths to show the effect on the distribution.

Keywords Reuse Distance, Time Distance, Random Access

1 Introduction

Reuse distance is the same as one of the stack distances by Mattson et al. [2]. The reuse distance of an access in an address trace is the number of distinct data elements referenced between the access and the nearest previous access that visits the same data element. The reuse distance is considered infinite if the data element is never visited before. For example, a trace is $a_1, b_1, c_1, b_2, d_1, c_2, a_2$. Assuming $a$, $b$, $c$, $d$ are four distinct data elements and the subscripts are the visit times. The reuse distances for this trace are $\infty, \infty, \infty, 1, \infty, 2, 3$. The time distance is the number of memory references between the two accesses to the same datum. For the above trace, the time distances are $\infty, \infty, \infty, 1, \infty, 2, 5$. Various algorithms have been developed to measure the reuse distance. See [1] for two approximate algorithms and references to earlier accurate techniques.

Random access is a frequently used assumption in performance modeling. In this paper, we show the proof that the reuse distances of random accesses follow a uniform distribution and use some experiments to check the convergence.

This problem is very similar to coupon collector problem - given a box with $n$ kinds of coupons, select one kind of coupon randomly and put it back each time, the expected times to get all kinds is $O(n \ln n)$ [3]. But we concern the number of kinds of coupon between collecting one kind of coupon repeatedly in this problem.
2 Reuse Distance Distribution

**Theorem 2.1.** The probability that a random access has reuse distance \(i\) (0 ≤ \(i\) ≤ \(n-1\)) is \(P(i) = \frac{1}{n}\), assuming each access has equal probability accessing one of \(n\) data elements.

**Proof.** We have the following formulas for \(P(i)\).

- For \(i = 0\),
  \[
  P(i) = \frac{1}{n}
  \]

- For \(1 ≤ i ≤ n - 1\),
  \[
  P(i) = \sum_{l=1}^{\infty} \left\{ \binom{n-1}{i} \sum_{l_1+l_2+\cdots+l_i=l \atop l_1 \geq 1, l_2 \geq 1, \ldots, l_i \geq 1} \left( \binom{l}{l_1, l_2, \ldots, l_i} \right) \left( \frac{1}{n} \right)^{l+1} \right\}
  \]

Now, we start to prove the theorem by induction.

- When \(i = 0\), we already have the fact that \(P(0) = \frac{1}{n}\).

- Assume when \(i \leq m\) (0 ≤ \(m\) ≤ \(n - 2\)), \(P(i) = \frac{1}{n}\).
\[
P(m+1) = \binom{n-1}{m+1} \sum_{l=m+1}^{\infty} \left\{ \sum_{l_1+l_2+\cdots+l_{m+1}=l} \binom{l}{l_1, l_2, \ldots, l_{m+1}} \left( \frac{1}{n} \right)^{l+1} \right\}
\]

\[
= \binom{n-1}{m+1} \sum_{l=0}^{\infty} \left\{ \sum_{l_1+l_2+\cdots+l_{m+1}=l} \binom{l}{l_1, l_2, \ldots, l_{m+1}} \left( \frac{1}{n} \right)^{l+1} \right\}
\]

\[
= \binom{n-1}{m+1} \sum_{l=0}^{\infty} \left\{ \sum_{l_1+l_2+\cdots+l_{m+1}=l} \binom{l}{l_1, l_2, \ldots, l_{m+1}} \left( \frac{1}{n} \right)^{l+1} \right\}
\]

\[
- \binom{n-1}{m+1} \sum_{j=m}^{\infty} \left\{ \sum_{l=0}^{\infty} \left( \frac{m+1}{l} \right) \sum_{l_1+l_2+\cdots+l_{m+1}=l} \binom{l}{l_1, l_2, \ldots, l_{m+1}} \left( \frac{1}{n} \right)^{l+1} \right\} \frac{1}{n} \]

\[
= \binom{n-1}{m+1} \sum_{l=0}^{\infty} \left\{ \sum_{l_1+l_2+\cdots+l_{m+1}=l} \binom{l}{l_1, l_2, \ldots, l_{m+1}} \left( \frac{1}{n} \right)^{l+1} \right\}
\]

\[
= \frac{1}{n-m-1} \binom{n-1}{m+1} \sum_{j=m}^{\infty} \left\{ \sum_{l=0}^{\infty} \left( \frac{m+1}{l} \right) \sum_{l_1+l_2+\cdots+l_{m+1}=l} \binom{l}{l_1, l_2, \ldots, l_{m+1}} \left( \frac{1}{n} \right)^{l+1} \right\} \frac{1}{n} \]

Because for \(1 \leq i \leq m,

\[
P(i) = \binom{n-1}{i} \sum_{l=i}^{\infty} \left\{ \sum_{l_1+l_2+\cdots+l_{i+1}=l} \binom{l}{l_1, l_2, \ldots, l_i} \left( \frac{1}{n} \right)^{l+1} \right\}
\]

\[
= \binom{n-1}{i} \sum_{l=0}^{\infty} \left\{ \sum_{l_1+l_2+\cdots+l_{i+1}=l} \binom{l}{l_1, l_2, \ldots, l_i} \left( \frac{1}{n} \right)^{l+1} \right\}
\]
\[ P(m+1) = \frac{1}{n-m-1} \binom{n-1}{m+1} - \binom{n-1}{m+1} \sum_{j=m}^{0} \binom{m+1}{j} \frac{P(j)}{\binom{n-1}{j}} \]
\[ = \frac{1}{n-m-1} \binom{n-1}{m+1} - \frac{1}{n} \sum_{j=0}^{m+1} \binom{n-1-j}{n-m-2} + \frac{1}{n} \]
\[ = \frac{1}{n} \sum_{j=0}^{m+1} \binom{n-1-j}{n-m-2} + \frac{1}{n} \]
\[ = \frac{1}{n} \]

From the above two parts, it is proved that \( P(i) = \frac{1}{n} \) (0 \( \leq i \leq n-1 \)). So the reuse distances are uniformly distributed from 0 to \( n-1 \).

\( \square \)

3 A Shorter Proof

William Thies, after hearing about our result, has devised a much shorter proof. Following is the sketch of his proof.

- Assume we are studying the reuse distance of an access to date element \( x \).
- All data elements including \( x \) will be referenced in the following infinite long trace eventually.
- Neglect all non-first occurrences for all data elements in the following infinite long trace.
- The following infinite long trace is filtered to a permutation of all data elements.
- \( x \) may be in any position of the permutation with equal possibility.

4 Experiments

We set up a simple experiment to check the convergence, that is, how many random references before the reuse distances show the expected uniform distribution. We use a group 128 data and test two different trace lengths, 65536 and 1048576.

Figures 1 shows the results from the two traces. The \( x \)-axis is reuse distance. The \( y \)-axis is the probability or the percentage references having a reuse distance. We omit the infinite reuse distance.

By comparing the reuse distance distributions in the two figures, we see that the distributions become more consistent with the theoretical result as the trace becomes longer. This shows the limitation of the assumption of the theorem that a trace is of infinite length.

5 Discussion and Future Work

In the future we will generalize this work to multiple groups, which may help us to understand the memory behavior for multi-threaded programs.
Figure 1: Reuse Distance Distribution

Acknowledgment  Chen Ding posed this interesting problem and checked the proof. Qi Ge provided some very valuable advice on the proof.

References

