Sorting (5)

Michael L. Littman

September 10th, 1998

1 SORTING

1.1 Telephone Book Publishing

GTE Directories is one of the world's largest telephone directory companies. It annually publishes or provides sales and other directory-related services for nearly 2,400 directory titles, with a total circulation of 80 million copies.
GTE publishes a telephone directory for Durham. There are approximately 100 thousand names listed in the directory. For the sake of human beings, it is helpful to publish those names in alphabetical order (Why?)
Obviously, this is done with the help of a computer.

1.2 Data Details

The data for each telephone is stored as a *record*, which consists of a telephone number, an address, and a name.
The name field acts as the *key*, in that we are interested in sorting according to the names.
Example:

- **Name:** Garrisom Al (key)
- **Address:** 2011 Evans Rd Creedmore
- **Phone:** 528-6263

1.3 Comparing Keys

Let "$\leq$" be a relation on keys that induces a total ordering.
That just means that any two keys can be compared with "less than or equal to."
For the phone-book problem, "less than or equal to" refers to lexicographic ordering. For other problems, it might mean "no shorter than" or "to the left of", etc.
For current purposes, we need not assume that that keys are numbers, but it is sometimes easier to think of them that way.

1.4 Formal Statement [Step 0]

Definition of a *sorting algorithm*:

- Input: A list of n objects $A[1], \ldots, A[n]$, and a total order relation "$\leq$" on objects.
- Here, $A[i]$ refers to the ith object in the list.
- Output: A permutation B of A such that
- $B[1] \leq B[2] \leq B[3] \leq \ldots \leq B[n-1] \leq B[n]$.

1.5 Permutations

It is important for B to be a permutation of A, meaning that there is a one-to-one mapping between objects in B and objects in A.
That is, we can't just set $B[i] = A[1]$ for all i, even though that would satisfy the properties that:

- Every object in B is an object in A,
- $B[1] \leq B[2] \leq B[3] \leq \ldots \leq B[n-1] \leq B[n]$

In particular, it fails to satisfy the property that:

- Every object in A is an object in B.

With this added constraint, the definition is complete.

1.6 Beyond Adjacency

Note that the definition is given in terms of adjacent elements only.
However, we can make the following inference.
Lemma: If B is sorted, $1 \leq i \leq j \leq n$ implies $B[i] \leq B[j]$.
That means that every object to the left of $B[j]$ is smaller than $B[j]$ (not just the adjacent one).

- 0 3 4 6 6 7 9 A
- 1 2 3 4 5 6 7 i

How would you prove this?

1.7 Keys

This formalization of sorting basically requires that we specify sorting algorithms in terms of comparisons and duplications of keys.
To deal with large records, we can sort the keys along with pointers to the entire record (so we don't have to keep moving the whole object around).
In some cases, the keys themselves are very large. In this case, both comparison and duplication could take substantial amounts of time. More sophisticated algorithms are needed in this case (especially if we assume that the data doesn't fit into main memory!).

1.8 Sorts of Sorts

An *in-place* sort reuses the space in the A array to return the sorted list. It uses only a constant amount of scratch space (i.e., an amount independent of n).
A *stable* sort has the property that if $A[i] = A[j]$ and $i < j$ with $B[k]$ correponding to $A[i]$ and $B[l]$ to $A[j]$, then $k < l$ (i.e., the order of tied elements is not changed).

1.9 Other Applications

In other applications:

- Web search engines like "excite" score all documents and return them in in order of the scores.
- Some text editors allow you to sort the paragraphs in a document.

In addition, many algorithms use sorting internally to help organize data:

- minimum spanning tree (Kruskal)
- convex hull (Graham Scan).

2 INSERTION SORT

2.1 Arrays

We will be introducing some more sophisticated algorithms today.
The newest thing we'll see is an array data structure.
We write $A[i]$ to mean the ith element in an array (matrix, vector, ordered list) A.
In analyses, we assume that array references (assignments and lookups) take unit time.

2.2 Swaps

```
void swap( vector<int> & data, int a, int b )
{
  int tmp = data[a];
  data[a] = data[b];
  data[b] = tmp;
}
```

2.3 Step 1: Develop Algorithm

Here is our first attempt at a sorting algorithm.

```
void insertionsort( vector<int> & data )
{
  int i, n;

  n = data.size();

  for( unsigned int j=1; j < n; j++ )
  {
    for(i = j - 1; i >= 0 && data[i] > data[i+1]; i--)
    {
      swap (data, i, i+1);
    }
  }
}
```

2.4 Explanation

The outer "for" loop at iteration j has data[0] through data[$j-1$] sorted.
The inner "for" loop slides the next item into its place.

2.5 Step 2: Prove Correctness

Does this actually work?
Proof by induction.
Lemma: After iteration j (1 through $n-1$), keys 0 through j are sorted.
Proof:

- Base case: When $j = 1$, the single key data[0] is sorted.
- Inductive step: Assume after iteration $j-1$, keys 0 through $j-1$ are sorted.

- Note that we swap keys j and $j-1$ if they are out of order. We then continue to move the smaller element down, until it is either first or the item below it is smaller than it. By the inductive hypothesis, the initial segment of the list is sorted. By the current argument, the new item is in place, and by the inductive hypothesis, the final segment of the list is sorted.

2.6 Step 3: Efficiency

An easy upper bound is obtained by bounding the two loops by n. This gives a running time of $O(n^2)$.
Is the worst-case running time $\Theta(n^2)$? That is, are there inputs for which running time is quadratic?
$\sum_{i=1}^{n-1} i = 1/2(n-1)n = \Theta(n^2)$.
Is the running time $\Theta(n^2)$? That is, is it quadratic for *all* inputs? What is a best-case input for insertion sort?
$\sum_{i=1}^{n-1} 1 = n-1 = \Theta(n)$.

2.7 Measures of Sortedness

Often it useful to talk about measures of list sortedness. Such a *distance* measure should assign a score of zero to a sorted list, and any unsorted list should score a positive amount. Some measures:

- Unsortedp: Is the list completely sorted? Score 1 if no.
- Items out of place: Score 1 for each position that contains an element other than what would be there in a sorted list. Maximum distance is n.
- Inversions: Score 1 for every pair that is out of order. Maximum distance is $n(n-1)/2$.
- Adjacent inversions: Score 1 for every adjacent pair that is out of order. Maximum distance is $n-1$.
- Insertion index: Find the longest increasing subsequence (LIS). Score n minus the length of the LIS. The same as asking, what's the smallest number of keys I can remove and the reinsert to get a sorted list? Maximum distance is $n-1$.

Example:

- **2 1 6 3 5 4 7**
- unsortedp: 1
- items out of place: 5.
- inversions: 1+0+3+0+1=5.
- adjacent inversions: 3.
- insertion index: 7-4=3.

2.8 Sorting with Small Insertion Indices

It is very common to apply sorting algorithms to lists with small insertion indices. Example? Insertion sort is nice because it runs fast for lists with small insertion indices.

- Given a list with insertion index k, how many times will insertion sort have to bubble an item into place?
- What's the maximum work that it takes to bubble a single item into place?
- This gives a $O(kn + n)$ bound. In the homework, you'll show that this bound is not tight.

2.9 An Assortment of Sorts

Several other sorts have similar asymptotic profiles to insertion sort, with different advantages and disadvantages.

- insertion sort: To sort $A[1] \dots A[n]$, recursively sort $A[1] \dots A[n-1]$. Then, bubble $A[n]$ downward until it stops. $T(n) = n + T(n-1)$. Worst case $\Theta(n^2)$, best case $\Theta(n)$ on sorted list.
- selection sort: To sort $A[1] \dots A[n]$, pick out the max in $\Theta(n)$. Move max to the end in one swap. Recurse on $A[1] \dots A[n-1]$. $T(n) = n + T(n-1)$. Running time is always $\Theta(n^2)$. Swaps $\Theta(n)$, which is asymptotically optimal.
- Bubble sort: Scan through the list over and over again swapping adjacent elements that are out of order. Like insertion sort only much slower.

3 DIVIDE-AND-CONQUER SORTING

3.1 Recurrences for Earlier Sorts

We can write recurrences for the previous sorts as follows:

- insertion sort: $T(n) = T(n-1) + n$
- Bubblesort, selection sort: $T(n) = n + T(n-1)$.

All basically focus on sorting a single element per iteration.

3.2 Divide and Conquer

Two choices:

- We can blindly break the problem into smaller lists, sort them, then do some work to patch things up after the fact (like insertion sort).
- Or we can do some work up front so that when the subproblems are solved, we are left with a sorted list (like selection sort).

But we need to break our lists in half: $T(n) = n + 2T(n/2)$.
By balancing, we get $T(n) = n \log n$.

3.3 Two Algorithms

Based on these insights, we can create the skeleton of two distinct algorithms.

- Break the list into two lists of size $n/2$. Sort each separately. Now, put the two sorted lists together in linear time.
- In linear time, split the list into two lists such that all the small elements are in one list and all the big elements are in the other. Once the two sublists are sorted recursively, the final result is sorted.

Recognize these algorithms?

3.4 Merge Sort

Structure: Break the list into two equal-size sublists. Recursively sort the two sublists. Merge them together to form a single sorted list.
Tricky part: How do you merge?
Insertion sort is a special case in which we merge a sorted list of size $n-1$ and one of size 1.
Idea: Look at the first element in each list. Remove the smaller one (by definition, an empty list is never the smaller one). Repeat. Animated example.
This involves constant work per element of the final list, i.e., linear time. (To be analyzed later.)
Awkwardness: Not "in place."

3.5 Quicksort

Structure: Separate list into the "lows" and "highs." Put the "lows" in the left section of the array and the "highs" in the right. Sort each recursively. Resulting list is definitely sorted.
Tricky part: How separate list?
Selection sort is a special case in which we define the "highs" to be a set of size 1 and the lows all the rest.

Idea: We choose an element in the list to be the "pivot." Easy to create two sublists from this, depending on which side of the pivot we're on. Animated example.
This involves constant work per element, because we simply check each one and then stow it away according to whether it's bigger or smaller than the pivot.
The partitioning procedure is especially elegant because it can be done in place.
Awkwardness: Hard to pick a good pivot fast.

3.6 Partitioning In Place

```
int partition( vector<int> & data, int start, int end, int pivot_position )
{
  int x = data[pivot_position];
  int i = start - 1;
  int j = end + 1;

  while( true )
  {
    do j--;
    while( data[j] > x );

    do i++;
    while( data[i] < x );

    if( i < j )
      swap( data, i, j );
    else
      return j;
  }
}
```

3.7 Explanation

Start at opposites sides of the segment of array we'd like to partition.
Move left pointer inwards as long as elements are less than pivot.
Move right pointer inwards as long as elements are greater than pivot.
Swap elements under left and right pointers and repeat.

3.8 Picking A Pivot

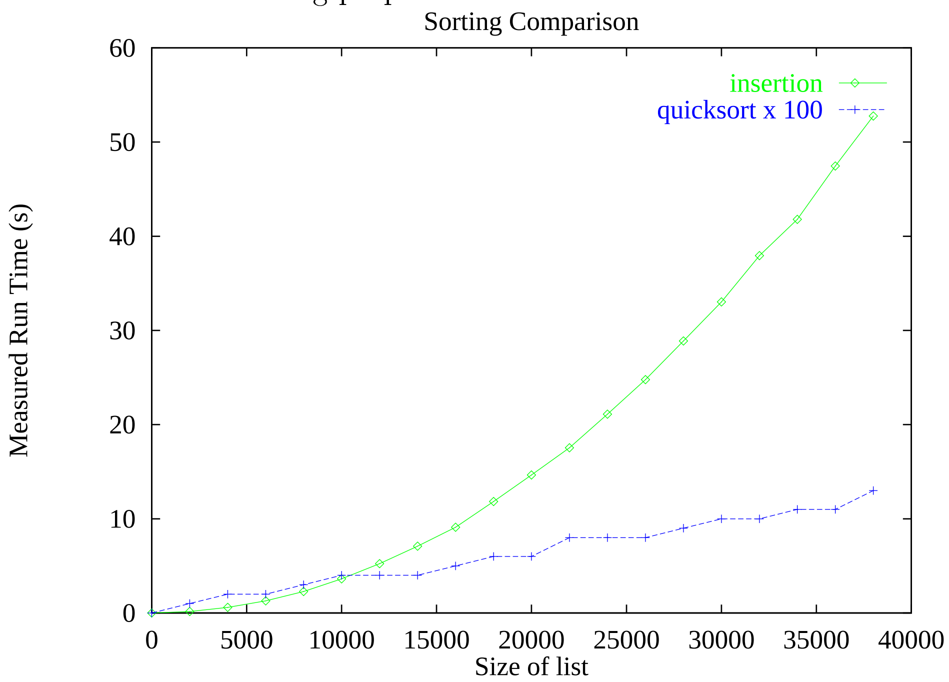
What kind of pivot do we want to pick?

- Max or min? No, then we get selection sort because one list will have $n - 1$ elements and the other 0.
- First element in list? Not good if list sorted or reverse sorted (get max or min)!

- Median of list? That would be perfect... $T(n) = n + 2T(n/2)$. Hard to find quickly. (Can be done in linear time, however.)
- Random element in list? Good chance of breaking the list into substantial chunks. Fast. But random... more complex analysis. (Next time.)
- Median of 3 randoms? More work, but more stable.

3.9 Empirical Comparison

Quicksort has better scaling properties *and* a better constant!



Note that the quicksort data is scaled up to make it visible on this plot!

3.10 A Consortium

There are some really nice demos of sorting algorithms on the web. Here are my favorites:

- High-level view of several sorting algorithms. Boy, quicksort is cool. Smaller set of algorithms.
- Nicer animations of how the algorithms work. Also provides mechanisms for collecting data to compare the variety of sorts provided.

4 SORTING IN LINEAR TIME

4.1 A Lower Bound

Several sorts are known that take $n \log n$ time (mergesort, quicksort, heapsort, and several others).
Is that as fast as one can go?
If we insist that sorting by done by comparison only, the answer is yes!

4.2 The Decision Tree

No matter how we sort, if we plan to put n items in order, we need to compare a sufficient number of items to uniquely identify the permutation.

- So, how many different permutations are there?
- Each comparison has the effect of ruling out some of the permutations. So, we can imagine that we start out with all permutations being possible, then making a comparison and splitting the set of possibilities in two.
- The entire sequence of comparisons for any given sorting algorithm can be thought of as a big tree, where each leaf of the tree contains no more than one permutation.
- The deepest path in this tree is the maximum number of comparisons we need to do to sort any list. We'll show that this is $\Omega(n \log n)$.

4.3 The Bound

Since a binary tree of height h has no more than 2^h leaves, we have: $2^h \geq n! \geq (n/2)^{n/2}$. The latter inequality comes from breaking up the factorial function:

$$n! = (n)(n-1)\cdots(n/2)(n/2-1)\cdots(2)(1) \geq (n/2)(n/2)\cdots(n/2)(1)\dots(1)(1) = (n/2)^{n/2}.$$

Taking logs and using their monotonicity property, we have $h \geq \log((n/2)^{n/2}) = n/2 \log(n) - n/2 = \Omega(n \log n)$.
So, the height of the tree is at least $\Omega(n \log n)$... with any fewer comparisons, we won't be able to sort.

4.4 Faster Sorts Via Bit Operations

The only way to beat the $n \log n$ bound is to use something other than comparisons in the sort.
Some well known sorts use various properties to achieve $O(n)$ running times: counting sort, radix sort, bucket sort.
I'll describe one that is an in-place sort that runs in time $O(nb)$ time sorting integers with b bits.

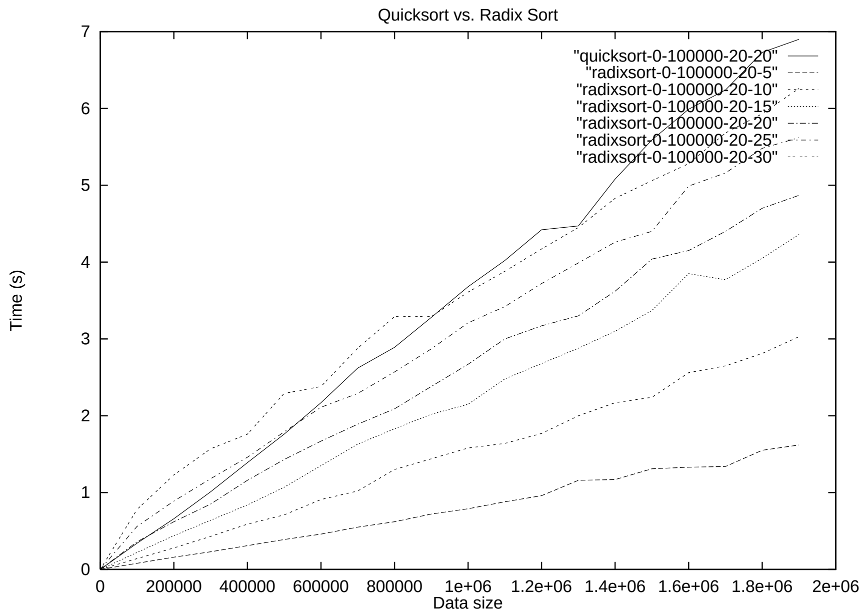
4.5 Forward Radix Sort

```
int partition_radix( vector<int> & data, int start, int end, int bit )
{
  int i = start - 1;
  int j = end + 1;

  while( true )
  {
    do j--;
    while( get_bit( data[j], bit ) == 1 && j >= start );

    do i++;
    while( get_bit( data[i], bit ) == 0 && i <= end );

    if( i < j )
      swap( data, i, j );
    else
      return j;
  }
}

void radixsort( vector<int> & data, int start, int end, int bit )
{
  if( start >= end ) return;
  if( bit == -1 ) return;

  int middle = partition_radix( data, start, end, bit );
  radixsort( data, start, middle, bit - 1 );
  radixsort( data, middle + 1, end, bit - 1 );
}
```

4.6 Discussion

This looks at the most significant bit, partitions the list into those with a zero there and those with a one. (Like quicksort.)
At this point, we can sort the two sublists separately.
Recurrence: $T(n, 0) = 0$, $T(n, b) = n + T(a, b-1) + T(n-a, n-1)$ (for some $1 \leq a \leq n$).
Can prove by induction that this gives $T(n) = bn$.

4.7 Empirical Results on Random Lists



5 SPARSE MATRICES

5.1 Matrices

An m by n matrix A contains nm elements in it. For $1 \leq i \leq m$, $1 \leq j \leq n$, $A[i,j]$ is the (i,j)th element of matrix A.
Matrices can be used to represent all kinds of things. $A[i,j]$ might be:

- the grade for student i on exam j;
- the driving distance from city i to city j;
- the number of times word i appears in encyclopedia article j;
- the probability of an elevator going from floor i to floor j.

5.2 Vectors and Matrix-Vector Product

A vector is just a $n \times 1$ matrix.
The *product* of an $m \times n$ matrix A and an $n \times 1$ vector x is an $m \times 1$ vector y where $y_i = \sum_{j=1}^{n} A[i,j]x[j]$. It is the "dot product" between the rows of A and x.
Matrix-vector products come up all over the place. For example:

- If A is the matrix of student exam grades and x is a vector of "weights" (what fraction of the final grade is assigned to each exam), then Ax is a vector of final scores for each student.

- If $A[i,j]$ is the number of times word j appears in article i and $x[j]$ is 1 if and only if a "query" contains term j, then Ax is a vector with one entry per article with the total number of query terms appearing in that article.

5.3 Transposition and Matrix-Matrix Products

The transpose of a m by n matrix A is a n by m matrix A^T such that $A^T[i,j] = A[j,i]$. We've exchanged rows and columns.
The product of two matrices A (m by n) and B (n by p) is a matrix C (m by p) where $C[i,j] = \sum_k A[i,k]B[k,j]$. Note that this is most often computed using a simple triply nested loop.
Matrix products come up a lot, too.

- If A is the matrix of floor-floor transition probabilities ($A[i,j]$ is the probability that the elevator will next be on floor j given that it is currently on floor i), then $AA = A^2$ is the matrix of two-step transition probabilities (the i,j entry is the probability that the elevator will be on floor j two stops from now given that it is currently on floor i).

5.4 Sparse Matrices

In many applications, we need to manipulate *sparse* matrices. These are matrices that consist of many zeros and very few non-zero elements.
For example, the word-document matrix for an encyclopedia is extremely sparse. Most words do not appear in most articles.
Instead of taking mn space to represent a matrix A with l non-zero elements, we can represent it using $\Theta(l)$ space using three vectors: for $1 \le i \le l$, $A_c[i]$ is the column of non-zero element i, $A_r[i]$ is its row, and $A_v[i]$ is its value. For ease of computation, we require that $A_c[i] \le A_c[j]$ if $i \le j$. Further, if $A_c[i] = A_c[j]$, then $A_r[i] \le A_r[j]$. This means that the non-zero elements are sorted by column, then by row within column.
Given an m by n matrix A, we can convert it to the sparse representation in time $\Theta(mn)$. (Simply read across the rows in order, one at a time, creating an entry in the sparse matrix format for each non-zero encountered.) Given a sparse matrix A with l non-zero elements, we can convert it to a standard (dense) matrix starting from a matrix of all zeros in $\Theta(l)$ time. (Simply assign $A[A_r[i], A_c[i]] = A_v[i]$ for each $1 \le i \le l$.)

5.5 Problems

How fast can we do each of these?

- Let A be a sparse m by n matrix with l non-zero elements. For $1 \le i \le m$, $1 \le j \le n$, compute $A[i,j]$.
- Let b be a dense 1 by m row vector. Let A be a sparse m by n matrix with l non-zero elements. Compute b times A.
- Same problem again, but let b be sparse with k non-zero elements.

- Let A be a sparse m by n matrix with l non-zero elements. Compute the transpose of A.
- Let A be a sparse m by n matrix with l non-zero elements. Let B be a sparse n by p matrix with k non-zero elements. Compute A times B.