ECS 222A: Assignment #1

Due on Tusday, January 13, 2015

Daniel Gusfield TR 4:40pm-6:00pm

Wenhao Wu

Contents

Problem 1

Given two strings L and S that have an equal number of occurrences of each specific character, we define $D(S_1, S_2)$ as the minimum number of transpositions of adjacent characters needed to convert S_1 into S_2. For example, $S_1 = CBADA$ can be converted into $S_2 = ABDCA$ using exactly four transpositions. Notice that all the transpositions are done on S_1.

Problem 1(a)

Assume that S_1 and S_2 each have exactly one occurrence of each character, for example $S_1 = ACBD$ and $S_2 = DCAB$. Develop an efficient algorithm to compute the number $D(S_1, S_2)$ given any input strings S_1 and S_2 that obey the stated assumption. Argue that your algorithm is correct (try to find a rigorous yet simple argument) and discuss how efficient it is in terms of the number of operations it does. (What are the primitive operations in your algorithm?)

Answer:

Lemma 1.1 *$D(S_1, S_2)$ equals to the Kendall tau distance between S_1 and S_2, i.e.*

$$\begin{aligned} D(S_1, S_2) &= K(S_1, S_2) \\ &= |\{(i,j) | i < j, (\tau_1[i] < \tau_1[j] \ AND \ \tau_2[i] > \tau_2[j]) \ OR \ (\tau_1[i] > \tau_1[j] \ AND \ \tau_2[i] < \tau_2[j])\}| \end{aligned}$$

where $\tau_1[i]$ and $\tau_2[i]$ denotes the index/ranking of character i in S_1 and S_2, respectively.

Proof Firstly, each transpositions of adjacent characters in S_1 can at most reduce $K(S_1, S_2)$ by 1, and when S_1 is eventually converted to S_2 with transpositions $K(S_1, S_2) = 0$. As a result, $K(S_1, S_2)$ is a lower bound of $D(S_1, S_2)$.
On the other hand, bubble sort uses exactly $K(S_1, S_2)$ adjacent transpositions to convert S_1 to S_2, therefore $K(S_1, S_2)$ is achievable. Consequently, $D(S_1, S_2) = K(S_1, S_2)$

A classical algorithm to compute $K(S_1, S_2)$ is merge-sort, has $\mathcal{O}(n \log n)$ complexity. Define index sequence Q such that

$$Q[k] = \tau_1[S_2[k]], \ k = 1, \ldots, n$$

the constructing of which has $\mathcal{O}(n)$ complexity. Then we can simply count the number of inverted pairs in Q using merge sort to compute $D(S_1, S_2)$.

Problem 1(b)

Develop an efficient algorithm to actually transform S_1 into S_2 using exactly $D(S_1, S_2)$ transpositions. Argue that your algorithm is correct and discuss how efficient it is in terms of the number of operations it does. The operations are the operations done by the algorithm, not the number of transpositions needed.

Answer: As stated in the proof to Lemma 1.1, we can bubble sort S_1 based on the pairwise ranking defined by Q, which uses exactly $D(S_1, S_2)$ transpositions to convert S_1 to S_2, since each transposition reduces $D(S_1, S_2)$ exactly by 1. It has $\mathcal{O}(n^2)$ complexity.

Problem 1(c)

Now explain how to handle the case of computing $D(S_1, S_2)$ when those strings may have more than one occurrence of any character. Assume that both strings have the same number of occurrences of any particular character. Hint: One approach is to find a simple reduction of this problem to the previous one. Of course, give a proof of correctness and an analysis of the number of operations needed.

Problem 1 [Problem 1(c)] continued on next page...

Answer:

Lemma 1.2 *The transpositions that achieve $D(S_1, S_2)$ does not change the order of multiple occurences of a same character. In other words, assume there are n_i occurences of character i in S_1. After S_1 being converted to S_2, each of this occurence appears in a new position denoted as $c_1^{(i)}, c_2^{(i)}, \ldots, c_{n_i}^{(i)}$, where the subscripts denotes the order of their original appearance in S_1, then a series of $D(S_1, S_2)$-achieving transpositions must satisfy*

$$c_1^{(i)} < c_2^{(i)} < \ldots < c_{n_i}^{(i)}$$

Proof If there exists $c_l^{(i)} > c_{l+1}^{(i)}$, then the l-th and the $l+1$-th occurence of character i must have been transposed somewhere, which is totally unnecessary. This is in contradict with the assumption of minimum number of transposition.

According to Lemma 1.2, we can redefine the index/ranking for each occurence of each character. Denote $\tau_1[i, m]$ as the index of the m-th occurence of character i in S_1, then the index sequence Q is also redefined as

$$Q[k_2[i, m]] = \tau_1[i, m],\ k = 1, \ldots, n$$

where $k_2[i, m]$ is the index of the m-th occurence of character i in S_2. Then counting the number of inverted pairs in Q using merge sort results in $D(S_1, S_2)$.
The algorithm to compute $D(S_1, S_2)$ and actually transform S_1 into S_2 is implemented in ***hw_1_1.py***.

Problem 2

The following algorithmic problem arose in the field of Sociology. You are given two strings, for example $L = ABCCBCD$ and $S = AQCBAD$. For each character X in the alphabet, define $M(X)$ as the minimum number of times X appears in either L or S. For example, $M(A) = 1$ and $M(Q) = 0$ in the example above.

The problem requires us to remove characters from L and S so that for each character X, the number of remaining occurrence of X in each of L and S is exactly $M(X)$. So far there is no real problem since we know exactly how many of each character must be removed from each string.

However, for some character(s) X there may be choices for which specific occurrences of X should be removed. Hence there are choices for what the resulting strings (call them S_1 and S_2) will be.

Now comes the problem: Among all possible ways to choose the required removals, we want to create resulting strings S_1 and S_2 to minimize $D(S_1, S_2)$. We call this the *MinDistRem problem*. For example, with L and S above, we can create $S_1 = ABCD$ and $S_2 = CBAD$ with $D(S_1, S_2) = 3$, or we could create $S_1 = S_2 = ACBD$ with $D(S_1, S_2) = 0$, and so we choose the latter as the solution to the *MinDistRem* problem.

Problem 2(a)

Solve the MinDistRem problem for $L = ACDQDCGFDERAE$ and $S = EECDACWERGARF$.

Answer: Denote $H_L[X]$ and $H_S[X]$ as the number of occurences of character X in string L and S, respectively. We have We first note that "Q" must be removed from L and "W" must be removed from S, while "A", "C", "F" and "G" should not be removed. 2 "D" must be removed from "L", 1 "E" must be removed from S and 1 "R" must be removed from S. We remove the characters that are surely to be

X	$H_L[X]$	$H_S[X]$	M[X]
A	2	2	2
C	2	2	2
D	3	1	1
E	2	3	2
F	1	1	1
G	1	1	1
Q	1	0	0
R	1	2	1
W	0	1	0

removed and highlight the characters that could be removed as

$$L = ACDDCGFDERAE$$
$$S = EECDACERGARF$$

There are a 3 ways of removing 2 "D"s from L and 2×2 ways to remove a "E" and "R" from S. We list the $D(S_1, S_2)$ for each possible pair of S_1 and S_2 in Table 1

Table 1: All possible $D(S_1, S_2)$ for exhaustive search.

S_2	$S_1 = ACCGFDERAE$	$S_1 = ACDCGFERAE$
$ECDACEGARF$	18	15
$ECDACERGAF$	18	15
$EECDACGARF$	22	19
$EECDACRGAF$	22	19

Using an exhaustive search, we have $S_1 = ACCGFDERAE$ and $S_2 = ECDACERGAF$ or $S2 = ECDACEGARF$.

Problem 2(b)

We would like to find an efficient algorithm for the MinDistRem problem. The following algorithm was proposed:

1. Let L be the longer string and S the shorter string. If the two strings are the same length, choose L and S arbitrarily.
2. For string L make a list (called LIST, duh!) of the characters (but not their positions) that will be removed. For each such character X include on LIST a number of copies of X equal to the number of occurrences of X that must be removed.
3. Set pointer p to 1 and pointers q and q' to 0.
4. Let X be the character in position p of S.
5. Scan L for the first occurrence (if any) of character X from position p forward in L. If such an X is found, set q' to the position of that found X.

6. If such an X was found in L, scan the characters from $q+1$ to $q'-1$ (inclusive) for any characters on the LIST. As each such character is found (if any are) remove it from L and remove one copy of it from the LIST.
7. Set q to q'.
8. Increment p by one.
9. Repeat steps 5 through 8 until p is at the end of S or until LIST is empty.
10. If LIST is not empty, scan L from its right end to find occurrences of characters on LIST. As each such character is found (if any are), remove it from L and remove one copy of it from LIST.
11. If sequence L is now shorter than sequence S, exchange the labels L and S, and repeat steps 2 to 10 on these two current strings L and S. Else terminate.
12. At termination, one the remaining string L is called S_1 and the remaining string S is called S_2.

Answer: This algorithm is implemented in ***hw_1_2.py***.

Problem 2(c)

Execute the above algorithm on $L = ABCCBCD$ and $S = ACBAD$ and then on $L = ACDQDCGFDERAE$ and $S = EECDACWERGARF$.

Answer: On $L = ABCCBCD$ and $S = ACBAD$ the glorithm results in $S_1 = S_2 = ACBD$. On $L = ACDQDCGFDERAE$ and $S = EECDACWERGARF$ the algorithm results in $S_1 = ECDACEGARF$ and $S_2 = ACCGFDERAE$.

Problem 2(d)

Argue that the algorithm always terminates within two executions of steps 2 to 10.

Answer: Steps 2 to 10 removes characters from L. Since step 10 ensures all characters in LIST are removed, and from the definition of LIST in step 2, the occurence of each character X in L is exactly $M[X]$ now.

Since S has not been altered yet, if S is in the same length as L, then the occurence of each character X in S is exactly $M[X]$ now and there is nothing to remove from it, the algotithm terminates in one pass of step 2 to 10. Otherwise the by swapping L and S the second pass will ensure so. Consequently, the algorithm always terminates within two executions of steps 2 to 10.

Problem 2(e)

Is it true that the algorithm always solves the MinDistRem problem? Justify.

Answer: Comparing the results from Problem 2(c) and Problem 2(a), we can see that this algorithm does not always solves the MinDistRem problem.

Problem 3

If you think the above algorithm does not solve the MinDistRem problem, can you propose an efficient algorithm that does always solve it? (this may be a very hard problem)

Answer: Too hard for me ...

Problem 4

A "prototein" is a binary string that we embed on a two-dimensional grid. A legal embedding must satisfy the following rules:

1. Each character in the string gets placed on one point of the grid.
2. Each point of the grid gets at most one character of the string placed on it.
3. Two adjacnet characters in the string must be placed on two points that are neighbors on the grid in either the horizontal or vertical direction, but not both. That is, two points across a diagonal are not neighbors. (Note that an interior point on the grid has four neighbor on the grid, and that an outside point has three neighbors, and that a corner point has only two neighbors).

Rules 1,2,3 mean that we are embedding the string onto the grid as a self-avoiding walk, without deforming the string. See the figure below. A contact is formed for every pair of 1's that are not adjacent in the string,

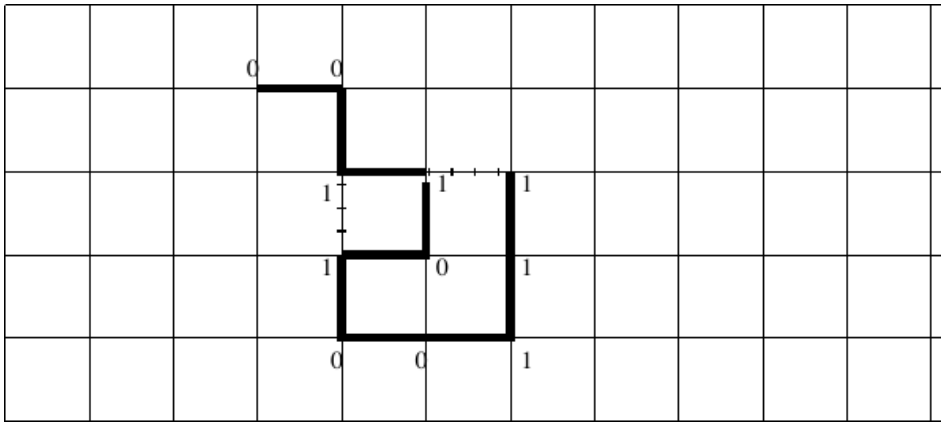

Figure 1: One possible embedding of the string 00110100111. The hatched lines show the contacts, two in this embedding. Can you find an embedding with more contacts?

but are placed on neighboring points in the grid. Such a pair of 1's is called a "contact" in the embedding.

Given an input string, the general problem is to find an embedding of the string on the grid so as to maximize the number of contacts.

Your Problems:

Problem 4(a)

Prove (that is give a clear explanation for) the following claim: For any string, and any legal embedding of the string, the characters in positions i and j in the string can form a contact only if $|ij|$ is odd. Try playing with some examples first to convince yourself that this is true, and then try to find a concise way to prove it.

Answer: We first color all grid points with 2 colors white and black so that the neighboring points have different color (as in a chess board).

Consider two points a and b on the grid, if they have the same color, then when b moves to one of its neighboring points, they have different colors; if they have different colors, then when b move to one of its

neighboring points, they have the same color. Consequently, two points of the same color must be separated by even steps and two points of different colors must be separated by odd steps. Here one step is defined as one move towards one of the neighboring points.

Due to the rule 3 the number of steps just defined equals to the distance between two characters in an embedded string. If characters on i and j forms a contact, their corresponding grid points must be neighbors so that they have different colors, therefore these two points must be separated by an odd number of steps. As a result $|i - j|$ is also odd.

Problem 4(b)

For any given string S, let $E(S)$ be the number of 1's in even positions in the string, and let $O(S)$ be the number of 1's in odd positions in the string. Let $C(S)$ be the minimum of $E(S)$ and $O(S)$. Prove that the number of contacts, in any legal embedding of S, cannot exceed $2(C(S) + 1)$.

Answer: According to problem 4(a), one character in a contact must be on an odd position and the other must be on an even position.

If $C(S) = O(S)$, we can count the number of contacts according to the 1's on the odd positions. For each of these 1's that is not at the beginning or end of the string, its corresponding grid points have at most 4 neighbors, but 2 of these 4 points must correspond to the 2 adjacent characters in the string so they can not form contact pairs. Consequently each 1 on the odd positions but not the beginning or end of the string can have at most 2 contacts. For the beginning 1 and the ending 1, a similar analysis applies and both of them have at most 3 contacts. As a result, the maximum number of contact is $2(C(S) + 1)$.

If $C(S) = E(S)$, we can count the number of contacts according to the 1's on the even positions. Similarly, each 1 that is not at the end of the string (it cannot be at the beginning!) can have at most 2 contacts. And the 1 at the end of the string (if any) can have at most 3 contacts. the maximum number of contact is $2C(S) + 1$.

In summary, the number of contacts in any legal embedding of S, cannot exceed $2(C(S) + 1)$.

Problem 4(c)

If we can invent the string S, as well as decide on how to embed it, we could get alot of contacts, but that is cheating. Still if we can invent a string S, and embed it, so that the ratio of the number of contacts to the number of 1's in the string is high, that would be a good challenge. Find a string S, and an embedding of it, that has a ratio of contacts to 1's of 7/6. Hint: you will need a string of length more than 25 (I think). So you will have to discover an idea, rather than just playing around.

Answer: One simple example is $S = 1001000000010010000000 1001$, and the embedding is shown in Figure 2. This embedding has exactly 7 contacts and 6 1's, so its ratio of contacts to 1's is 7/6.

Problem 4(d)

Prove that, over all possible strings and embeddings of those strings, the highest possible ratio of contacts to 1's is 7/6. Hint: the handshake lemma from graph theory (remember graphs from cs100 or cs20?) is helpful here. The rest is just case analysis.

Answer: Consider only the 1's in the string. According to how many neighboring 1's they have in the embedding, they can be categorized as follows:

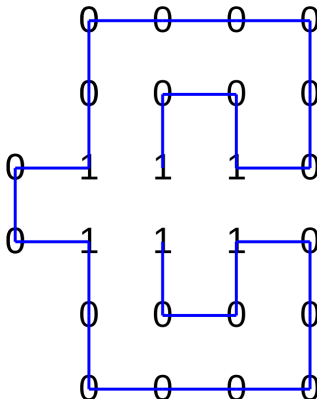

Figure 2: Embedding of $S = 1001000000010010000000 1001$.

1. The 1 with 4 neighboring 1's:
   - If this 1 is at the end or beginning of the string, it can form at most 3 contacts.
   - If this 1 is in the middle of the string, it can form at most 2 contacts.
2. The 1 with 3 neighboring 1's:
   - If this 1 is at the end or beginning of the string, it can form at most 3 contacts.
   - If this 1 is in the middle of the string, it can form at most 2 contacts.
3. The 1 with 2 neighboring 1's, it can form at most 2 contacts.
4. The 1 with 1 neighboring 1's, it can form at most 1 contacts.
5. The 1 with 0 neighboring 1's, it can form 0 contact.

According to the proof of the handshake lemma, we can count the total number of contacts as the summation of the number of contacts each 1 forms divided by 2. Then we sould like to prove that the average number of contacts each 1 can form is at most 7/3.

From the above categorization, we can see that if neither the end nor beginning of the string is a 1 with 3 neighboring 1's or 4 neighboring 1's, each 1 can form at most $2 < 7/3$ contacts.

If there is a 1 either at the begining or the end point of the string have 4 or 3 neighboring 1's, we would like to have as few 1's that is in the middle of the string as possible in order to make the average number of contacts each 1 can form larger, since each of them can form at most 2 contacts. According to the number of neighboring 1's that the beginning and ending 1's have, we can count the minimum number of middle 1's in the string and therefore compute the maximum possible ratio of contacts to 1's. The results are shown in Table

beginning of the string	end of the string	min. # 1's in the middle	max. ratio of contacts to 1's
1 with 4 neighbor 1's	1 with 4 neighbor 1's	6	$\frac{6\cdot 2+2\cdot 3}{6+2} =$
1 with 4 neighbor 1's	1 with 3 neighbor 1's	5	$\frac{5\cdot 2+2\cdot 3}{5+2} =$
1 with 4 neighbor 1's	1 with 2 neighbor 1's		
1 with 4 neighbor 1's	1 with 1 neighbor 1's		
1 with 4 neighbor 1's	1 with 0 neighbor 1's		
1 with 4 neighbor 1's	0		
1 with 3 neighbor 1's	1 with 3 neighbor 1's		
1 with 3 neighbor 1's	1 with 2 neighbor 1's		
1 with 3 neighbor 1's	1 with 1 neighbor 1's		
1 with 3 neighbor 1's	1 with 0 neighbor 1's		
1 with 3 neighbor 1's	0		