

Exam Questions

Approximation: Theory and Algorithms

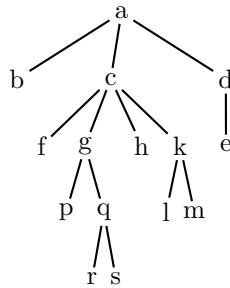
Nikolaus Augsten

Exam Date: June 23, 2009

1. Explain the framework for approximate matching. Explain techniques for matching and search space reduction by example. (*Unit 1*)
2. Compute the edit distance between two strings (size n and m) using the algorithm with $O(nm)$ time and $O(m)$ space complexity. Explain the algorithm. (*Unit 2*)
3. Compute the edit distance between the strings 'peter' and 'petre' and explain the algorithm. Use the matrix produced by the algorithm to derive the shortest edit scripts and represent them with the gap representation. (*Unit 2*)
4. Search the pattern 'abba' in the text 'ablaba' allowing 1 error and explain the algorithm. Use the matrix produced by the algorithm to find all possible matches and represent them with the gap representation. (*Unit 2*)
5. Write an approximate join query for the string attributes of the following tables A and B . Use the string edit distance with threshold $k = 1$. Include length filtering, count filtering, and position filtering into your query. How is the query evaluated? (*Unit 3*)

A		B	
id	$name$	id	$name$
1	Joe	6	Jeo
2	Peter	9	Peters

6. Represent the following tree in a relation using the adjacency lists encoding (interval encoding, dewey encoding). Insert node x as the 2^{nd} child of node c and move two children to the new node. How do you need to update your encoding? (*Unit 4-5*)



7. Explain the tree edit distance algorithm by Zhang-Shasha using the following example trees. (Unit 6-8)



8. Show time and space complexity of the tree edit distance algorithm by Zhang and Shasha. (Unit 9)
9. Show that the string distance between the preorder (postorder) traversals of two trees is a lower bound for the tree edit distance between them. Why is the traversal string distance a lower bound and not the exact tree distance? How are the two lower bounds (preorder and postorder distance) combined into one single lower bound? (Unit 9)
10. Consider the tree sets $S_1 = \{T_1, T_2, T_3, T_4\}$ and $S_2 = \{T_5, T_6, T_7, T_8\}$. Show the approximate join (threshold $\tau = 2$) based on the tree edit distance with and without a reference set. Choose a single tree for the reference set. The trees are shown in Figure 1, the distances between all pairs of trees are given in Table 1. (Unit 10)
11. Explain the triangle inequality and its application to approximate joins with well separated clusters. (Unit 10)
12. Explain Yang's algorithm for the binary branch distance between two trees by example and show that the binary branch distance is a lower bound for the tree edit distance. (Unit 11)
13. Explain the fanout weighted tree edit distance and the pq-gram distance between two trees by example. (Unit 11)
14. Compute the windowed pq-gram distance between the unordered trees in Figure 2 ($p = 2, q = 2, w = 3$) and explain the algorithm. (Unit 12)

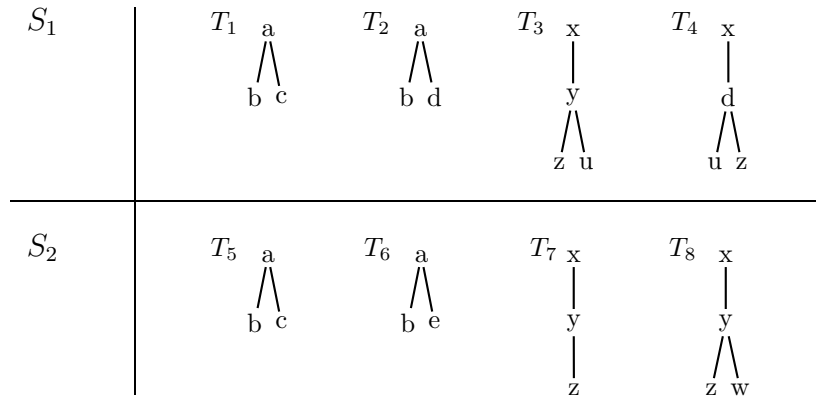


Figure 1: Trees for Question 10.

	T_1	T_2	T_3	T_4	T_5	T_6	T_7	T_8
T_1	0	1	4	4	0	1	4	4
T_2		0	4	4	1	1	4	4
T_3			0	3	4	4	1	1
T_4				0	4	4	2	3
T_5					0	1	4	4
T_6						0	4	4
T_7							0	1
T_8								0

Table 1: Distances between All Tree Pairs in Question 10.

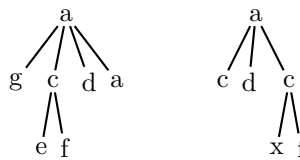


Figure 2: Trees for Question 14.

15. Explain by example, why the edit distance between ordered trees can not be used in combination with tree sorting to compute the distance between unordered trees? Why is this approach possible for windowed pq -grams?
(Unit 12)