Seminar on Advanced topics in data structures

Fall 2004/2005

Haim Kaplan
phone: 03-6408814
email: haimk@cs.tau.ac.il

October 19, 2004

We shall mainly focus on algorithms and data structures for processing strings. We will cover classical stuff such as
1) The algorithm of Boyer-Moore for classical exact pattern matching.
2) Suffix trees and suffix arrays, their construction, and applications.
Then we will go through more recent results such as pattern matching with errors, and overlap matching.

The last two lectures will cover a different data structure to maintain all consecutive ones ordering of 0/1 matrices.

How to get in: To participate in the seminar you must choose a paper and register it with me by Thursday (21/10) of this week. Most of the papers are available online through links from the course web site. You can also get hard copies either by coming to my office or at the library. Once you decide that you want to participate and picked a paper please come to register the paper at my office (its also possible to do that by phone or email if you cannot come in person). Please note that as soon as you register your chances to get in with the paper that you want are higher, in particular if there are a lot of students in class. Papers in the list below are ordered more or less in the order in which I would like them to be presented so you can estimate the approximate week of your presentation.

Once registered you must do the presentation you committed to!

Your grade: You'd have to attend all meetings and actively participate in the discussion so as to make sure you understand the material. The papers are strongly related. You should follow the material covered before your presentation and connect your presentation to it. (That does not necessarily mean that you do not have to repeat crucial definitions etc.) You are encouraged to ask questions and you get credit for it. To encourage you to ask everything you like I guarantee not to take any credit from the speaker even in case that he or she does not answer well. So there is absolutely no way you can hurt your friend by asking, and you may hurt yourself if you refrain from asking.

Participation worth up to 15% of your final mark. The other 85% will be determined by the quality of your presentation.

Guidelines for preparing your presentation:

1. All papers (or book chapters) on the list are mostly theoretical and you may have to spend some time in order to understand your paper well. Good understanding of the paper is the key for a good presentation. So start working early and try to gain a good understanding of the results.

2. Some of the papers on the list are related to others. So attending the other lectures would certainly help you when you work on your paper.

3. You are not restricted to use any particular media for your presentation. Clear conventional blackboard talk certainly can gain maximum credit. Remember that sometimes a picture worth 1000 words. Media such as slides or computer slides is particularly useful when you want to show pictures and animate operations since doing that on the board may take too much time. Your time is limited so you have to use it carefully.

4. Sometimes with the time you have you will not be able to cover everything that is written in your paper. Depending on how crowded the seminar would be, it may not be possible to give you extra time. So plan the talk carefully, you need not cover everything in the paper but what you cover should be transfered clearly. Don't cover stuff that is in your paper but has already been covered by previous speakers.

   If you think that you need less time than what you have since the paper is simple or short, please let me know as soon as you estimate it.

Lectures:

1. The algorithm of Boyer-Moore and its analysis (2.2,3.1,3.2 in [4]).

2. Linear time construction of suffix trees. (Chapter 6 in [4]) Present either one or two of the algorithms.

3. Constant-Time lowest common ancestor retrieval [8].

4. Selected applications of suffix trees (Based on Chapters 7 and 9 of [4]). Talk about repeatitions ,k-mismatch, tandem repeats, Ziv-Lempel, and others.

5. Suffix arrays [9, 7]

6. Seminumerical string matching (Chapter 4 of [4]).

7. pattern matching with k-mismatches [2]

8. overlap matching [1]

9. Compressed suffix arrays and suffix trees with applications to text indexing and string matching [3]

10. Burrows-Wheeler transform [10]

11. PC trees and circular-ones arrangements [6]

12. PQ tree PC tree and planarity testing [5]

References

[1] A. Amir, R. Cole, R. Hariharan, M. Lewenstein, and E. Porat. Overlap matching. *Inf. Comput.*, 181(1):57–74, 2003.

[2] A. Amir, M. Lewenstein, and E. Porat. Faster algorithms for string matching with k mismatches. *J. of Algorithms*, 50(2):257–275, 2004.

[3] R. Grossi and J. S. Vitter. Compressed suffix arrays and suffix trees with applications to text indexing and string matching (extended abstract). In *Proceedings of the thirty-second annual ACM symposium on Theory of computing*, pages 397–406. ACM Press, 2000.

[4] Dan Gusfield. *Algorithms on Strings, Trees and Graphs.* Cambridge Univ. Press, Cambridge, 1997.

[5] W. L. Hsu and R. McConnell. *PQ trees, PC Trees, and Planarity*, chapter 37. Handbook of Data Structures and Applications, D. Mehta and S. Sahni editors. CRC press. to appear.

[6] Wen-Lian Hsu and Ross M. McConnell. Pc trees and circular-ones arrangements. *Theor. Comput. Sci.*, 296(1):99–116, 2003.

[7] J. Kärkkäinen and P. Sanders. Simple linear work suffix array construction. In *Proceedings of the thirty-second annual ACM symposium on Theory of computing*, volume 2719 of *LNCS*, pages 943–955. springer, 2003.

[8] M. Farach-Colton M. A. Bender. The LCA problem revisted. In 4^{th} *LATIN*, pages 88–94, 2000.

[9] U. Manber and G. Myers. Suffix arrays: a new method for on-line string searches. *SIAM Journal on Computing*, 22(5):935–948, 1993.

[10] Giovanni Manzini. An analysis of the burrows wheeler transform. *J. ACM*, 48(3):407–430, 2001.