Examiner's Report, Ph. D. Thesis
Martin Senft
Faculty of Mathematics & Physics
Charles University in Prague:
"Suffix Graphs & Lossless Data Compression"

July 6, 2013

Overall Assessment

The thesis provides very thorough coverage of the suffix tree data structure and its variants (suffix trie, DAWG, and CDAWG), unified by the concept of a suffix graph. It describes a suffix graph construction algorithm with variants that applies to all four structures. In particular, for the sliding window used for data compression, the candidate shows that sliding the suffix tree can be accomplished in constant time, an important new result. Thus, using these insights, it becomes possible to design new data compression methods, based on different sliding strategies; at least one of these methods provides significantly better compression on important test cases.

The thesis provides the most thorough treatment of suffix graph construction — and use for data compression — that I have ever read. The view taken of suffix graphs is quite original, several open problems are solved, and the new approach taken by the candidate is likely to lead to still other new insights in future. The candidate clearly demonstrates a talent for creative scientific research. There is no doubt that this thesis would be accepted for the Ph.D. degree at McMaster University.

A nontrivial precondition for doing research in Computer Science is capability in English. In this context, the candidate is very well prepared:

although there is perhaps not a single sentence in the thesis that I, as a native English speaker, would not wish to change, nevertheless the thesis is extremely well written: it is well organized, and its author succeeds in making his meaning absolutely clear even though English is not his native language.

Suggestions for Improvement

- On pages 3–5 I thought that the basic concepts and definitions would be clearer if examples were provided of some of them at least. In general, there is a shortage of illustrative figures in the thesis.

- At the end of Chapters 3–5 I thought the author would do himself (and the reader) a service by providing in point form a summary of the main results/insights achieved within the current chapter.

- On page 32 under "Reverse Edges" the author says that "from now on, the discussion is limited to suffix trees only". First, this is not true, since other suffix structures are considered in later chapters; second, the new context should be made clear in headings — for example, Section 3.2.2 might become "Experimental Evaluation (Suffix Trees)".

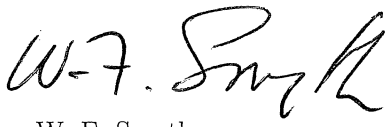
- Also under "Reverse Edges", the author mentions sliding for the first time (I believe), but without any definition. Indeed, as far as I could determine, sliding was never defined in a precise way. Also, for the definition of such a thing, a diagram is helpful.

- On page 35 I found the definition of "move-to-front" rather skimpy. There is much to be said about "move-to-front" and other similar heuristics. The author showed no awareness of the classic paper by Bentley, Sleator, Tarjan & Wie (1986), "A locally adaptive data compression scheme". This and other important contributions can be found at

  http://www.data-compression.info/Algorithms/MTF/

- On page 37 it would be helpful to define the IPM more precisely.

- On page 45 sliding is implemented before it is defined or discussed (see Chapter 4)
- On page 49 sliding is not defined; furthermore, pseudocode does not constitute a definition of "sliding operations". Here again a diagram would be also be helpful.
- Again on page 52 ff. the discussion would be greatly helped by diagrams.
- Unwanted blank space on pages 72 & 73.
- On page 88 especially an itemized statement of conclusions to be drawn from the tests of Chapter 5 is essential.



W. F, Smyth