Effective Virus Scanning Algorithms [6]

Zoltán Hornák and Endre Selényi

Large number of computer viruses and their world-wide spreading is an interesting but very harmful factor in today's personal computer systems. Most anti-viral techniques used today are based on virus searching methods, that is they can detect and identify only those known viruses which were previously caught by anti-viral specialists and the specific virus detection codes were added to the anti-virus software. There are many methods for detecting and identifying viruses. Among these methods virus scanning is the mostly used and the most effective in practice.

Because of the large number of computer viruses and the continuously increasing size of the area which should be scanned, only the most effective scanning algorithms can be applied in the practice. On gigabyte-sized hard disks scanning of some ten thousand viruses with a brute force algorithm would take more than 1013 comparisons, which cannot be performed on a PC in reasonable time.

In contract with the traditional algorithm theory, where the behaviour of algorithms is analysed in the infinite, the discussed problem deals with the fastest solution of a specific, limited task. The best algorithm found uses the techniques of precondition, hashing and also the consideration of statistical probabilities. Although the resulting techniques are specific for virus scanning they can be used in other scanning problems as well.

The problem of virus scanning is defined as a search for occurrences of sequences in an area. A sequence is a string of bytes extended by special so called joker characters including ?, which matches any byte, +(number), which matches any (number) bytes, *(number), which matches any 0 to (number) bytes in length and [<alternative 1>, <alternative2>], which matches <alternative1> or <alternative2> sequences respectively.

For an effective search for occurrences of many sequences some candidate solutions were examined from speed to memory requirement points of view. All of the proposed solutions were based on parallel search, and used hashing technique. Hashing was combined with the optimised step-back precondition technique to improve speed, that is after a byte-stream was checked the processing did not return to the beginning of the byte-stream, but to an optimal position where an occurrence of a sequence was possible. This solution was proved to be very fast, but required a huge amount of memory.

Possible improvement to the above hashing algorithms is the reordering of the sequences to make the optimal step-back technique more effective. It turned out that the reordering of sequences is most effective if we also consider the statistical probabilities of byte and word occurrences in the scanned area.

Finally a mathematical model was developed to formalise the requirements toward an optimal sequence reordering. Experiences of the implementation of the algorithm showed that with realistic virus sequences this solution is so effective that the average number of comparisons is somewhere around the one sixth of the length of the scanned area, which is significantly better than the results of single hashing algorithms.

[6]This research was funded by the Hungarian Scientific Research Fund (F026251)