

# Digital Documents in Educational Environment: Misuse, Appropriation, and Detection Issues

Krisztián Monostori, Arkady Zaslavsky, Heinz Schmidt

School of Computer Science and Software Engineering Monash University, Melbourne

900 Dandenong Road, Caulfield East, 3145 Australia

Tel:+61-3-9903-{1410,2479,2332} Fax:+61-3-9903-1077

E-mail: {krisztian.monostori, arkady.zaslavsky,  
heinz.schmidt}@infotech.monash.edu.au

With the proliferation of the Internet, research papers as well as other digital documents are easily available on the Internet. This is a great help for bona fide researchers but users can also misuse these documents. It is tempting for students to substitute genuine creativity by simply copying ideas and unaltered text from other papers.

This poster describes how we can tackle this problem and how we can catch students cheating. Available systems, including SCAM, Koala, the “shingling” approach, and the sif tool, use fairly similar approach. They divide the text into chunks and an index is created on chunks. A suspicious document is compared to the index and overlapping chunks are defined. Systems differ in chunking primitives and the number of chunks stored, and these parameters have an impact on the performance of these systems. Accuracy, space-efficiency, and speed of comparison are compared regarding these systems.

It is obvious that the more chunks we store the more space we need and the accuracy also increases with the number of chunks kept in the index. Of course, storing more chunks means more disk space and searching a larger index is also more time-consuming. If we decrease the number of chunks to be stored we have to pay the penalty of losing on accuracy. Two problems are false positives and false negatives. False positives are those documents that the system reports to have overlap with the suspicious document, though they do not overlap. False negatives are documents that overlap with the original document but they

are not reported by the system. False negatives can be eliminated by storing more chunks but as we have already discussed we have to limit the number of chunks to be stored because we want to index a large number of documents, i.e. the Internet. False positives are reported because hash values are calculated on chunks. Rather than storing the chunks themselves we store hash values that represent chunks. Not only do we expect false positives because hash functions can produce the same value for different chunks but also because the number of possible chunks in Internet-documents outnumber the available number of hash values.

There are two ways to reduce index-space. We can either reduce the number of chunks to be kept, which increases the chance of false negatives, or we can reduce the size of the hash value we calculate on each chunk, which increases the chance of false positives. False negatives are harder to handle because we have already missed potential documents. We propose a method that is able to eliminate false positives from a given set of documents. The comparison is completed in two phases. In the first phase we define candidate documents using the aforementioned methods and the second stage eliminates false positives. Our algorithm for eliminating false positives uses a suffix tree built on the suspicious document to compare candidate documents and eliminate accidental matches. Comparison of the chunking methods and our algorithm are presented in this poster.

**Monostori K., Zaslavsky A., Schmidt H. Digital Documents in Educational Environment: Misuse, Appropriation, and Detection Issues. *Fourth Australasian Computing Education Conference, 4 - 6 December 2000, Monash University, Melbourne. p.254, ACM Press***