Towards the Creation of a Robust Search Index for Digitalized Documents

László Kovács[1], Máté Pataki[1], Tamás Füzessy[2] and Zoltán Tóth[1]

[1]MTA SZTAKI, H-1111 Budapest, Lágymányosi u. 11.
laszlo.kovacs@sztaki.hu, pataki.mate@sztaki.hu, zoltan.toth@sztaki.hu
[2] FreeSoft Rt., H-1117 Budapest, Neumann János u. 1/C
tfuzessy@freesoft.hu

The simultaneous support of electronic and paper-based document handling is a natural demand of current filing and document management systems. To support the better management of search and retrieval functions and to reduce the high costs of digitizing, the Department of Distributed Systems of SZTAKI analysed the different kinds of error that emerged during the digitization process of Hungarian documents, and examined how these errors affect the searchability of the digitized items. For this reason, a testbed was set up that was suitable for the automatic analysis of digitized texts in a large corpus, and the conclusions and statistics obtained from the analysis were employed in the development of new content management products. The primary beneficiaries of these are civil service and higher-education bodies.

Today the realization of the ’almost paperless office’ can be achieved via post-digitization, or more precisely via scanning and OCR, as a huge number of documents still need to be digitized.

For various reasons, errors may occur during the digitization process; in seeking to achieve the highest quality for full text search capabilities, accuracy is thus an important issue. Therefore, the application of a search engine with high fault tolerance would make texts more suitable for search and retrieval purposes and would enhance their usability in practice while considerably reducing the costs of digitizing – primarily because post-processing human intervention to make corrections would be unnecessary.

The primary goal of the project was to build a metric for the errors introduced during the OCR process, particularly for those resulting in the loss or alteration of characters or accents, and to build a robust search index for digital repositories containing automatically digitized, error-prone documents.

Testbed for the evaluation of digitalization error types

Our testbed consisted of a large corpus containing Hungarian documents in various formats (rtf, txt, pdf and doc), digitizing software capable of character recognition from digital image formats, and a branch of self-developed utilities. The documents were converted to images, different kinds of noise were artificially generated over them (coffee-patches, traces of plying, noise), and they were then sent to the digitizing application. This resulted in digital, textual documents that could be compared with the originals.

The comparison process took place in two steps. First, a manual comparison was conducted for a small number of documents to identify categories of error types. An automatic comparison then took place for the whole corpus. Based on the results of the manual comparison, we evaluated our automatic method and generated different statistics to tune the categorization of the error types.

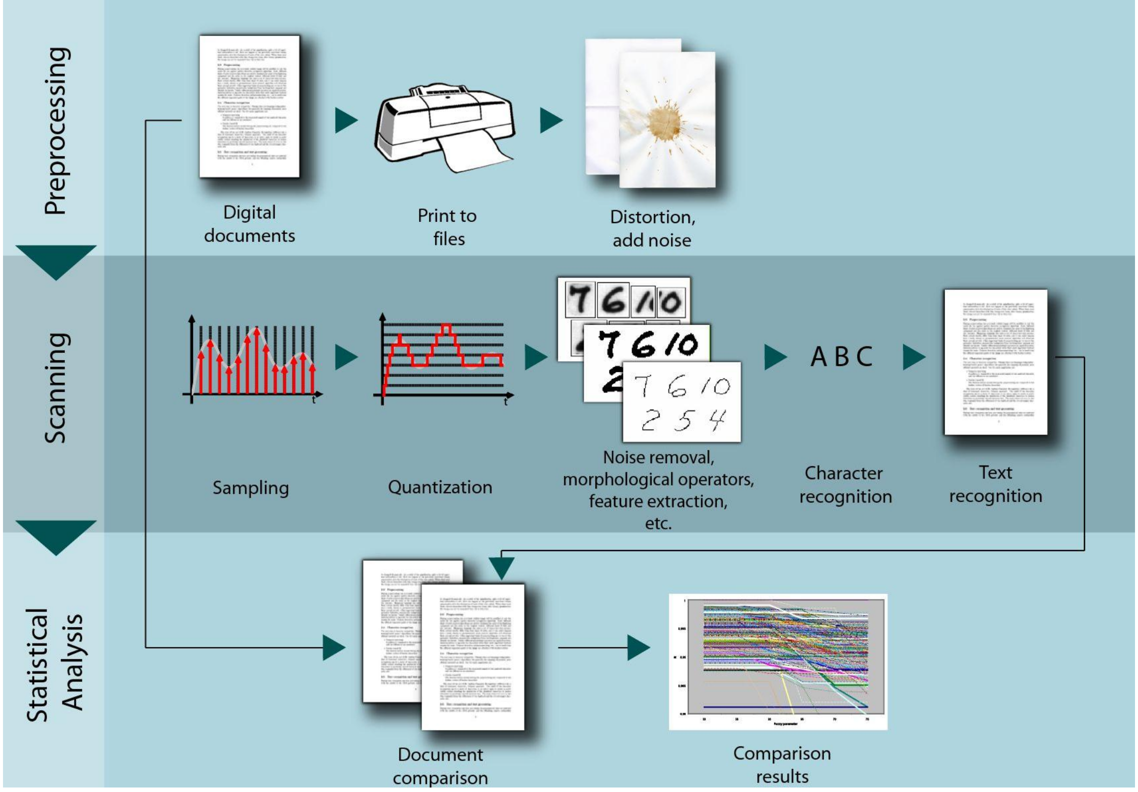

Figure 1: Architecture of the testbed.

Actual findings

As a result, another corpus was built containing words that had been altered to other words and typical accented characters that were recognized as other characters or character series. Though the digitization accuracy rate was quite high (around 95%, which is the expected value also mentioned in the literature), typical character/word changes that could strongly affect the searchability of the output document could still be detected. For example, the most common digitizing error concerns the letter ‘i’. Substitutions include the characters ‘j’, ‘l’, and ‘1’, which is not surprising as even to a human observer they may appear similar. In this case therefore, the creation of an index mapped to the letter ’i’ might present a solution to search-related problems.

Another example is the letter ‘m’, which was often recognized as ‘rn’ (‘r’ followed by ‘n’). Hence, when searching for words containing the letter ‘m’, one could also search for the same word having the ‘m’ replaced with ‘rn’. In reverse, this method is employed by spammers to obfuscate dictionary-based spam filters.

Further, our results confirm the hypothesis that errors related to accented characters like é, á, ő, and ö occur quite often. For example, the character ‘o’ has three accented variants in the Hungarian language (ö, ő, ó); together with the capital equivalents, this makes eight different but barely distinguishable characters for the OCR software. Even during post-processing, it is hard to tell which variant is the correct one, as there are many meaningful word-pairs that differ only in a single accent (eg kor, kór, kör). Complete statistics were gathered for the most common accented character identification errors.

The fault-tolerant search algorithm that was developed based on these findings has been integrated into the new versions of the Contentum content management product, and may also be used for further collaboration in European projects related to data repositories. In addition, and along with the list of the most common character substitutions, the analysis and the algorithm may provide a good basis in the future for building a robust search index for digital repositories comprising digitized documents.

Links:
Meta-Contentum R&D project (FreeSoft, "A meta-contentum k+f projekt.")
http://www.contentum.hu/hu/news/meta-contentum-kf
Department of Distributed Systems, MTA SZTAKI
http://dsd.sztaki.hu

Please contact:
László Kovács
SZTAKI
Tel: +36 1 279 6212
E-mail: laszlo.kovacs@sztaki.hu