Masakazu Suzuki

Mathematical Formulae Recognition and Logical Structure Analysis of Mathematical Papers

In: Petr Sojka (ed.): Towards a Digital Mathematics Library. Paris, France, July 7-8th, 2010. Masaryk University Press, Brno, Czech Republic, 2010. pp. 7--7.

Persistent URL: http://dml.cz/dmlcz/702568

Mathematical Formulae Recognition and Logical Structure Analysis of Mathematical Papers

Extended Abstract

Masakazu Suzuki

Kyushu University, Kyushu, Japan
suzuki@math.kyushu-u.ac.jp

Abstract. In most cases the current on-line journals in mathematics are supplied in the form of PDF with print images of papers in the front and OCR'ed hidden texts behind to provide with search facilily using key words. The embedded hidden texts usually does not include good information about mathematical formulae in the papers.
We can say that, for the future development of DML, it is desirable to include, in the digitised journals, more structured information of the content of mathematical papers, e.g. tag information to indicate logical structure of papers such as headings of sections, definitions, theorems, lemmas, etc., together with mathematical formulae structures included.
In the talk, I will present the current stage of our technology to extract such information from the scanned images in the retro-digitised mathematical papers. Mechanically-prepared new journals in the form of PDF are also the target of our research since it is not an easy task to get uniform structure description of mathematical formulae for example from the original LaTeX source with various styles and macro commands depending on authors.
Although there are many methods presented in literature to recognize mathematical formulae, very few applications appeared to do this task in practical sense. One of the major problem in the development of math OCR is to avoid fatal effects caused by mis-recognition and mis-segmentation of characters and symbols. In the talk, I will explain first the method we took to overcome this difficulty. Some demonstration of our software InftyReader to recognize mathematical documents will also be given in the lecture. Secondly, as a better approach to recognize a large number of pages like the case of DML, our adaptive method to improve the recognition rates of characters/symbols, mathematical formulae structures and logical structures of articles will also be presented.

Petr Sojka (editor): DML 2010, Towards a Digital Mathematics Library, pp. 7–8.
 ISBN 978-80-210-5242-0