Caractérisation d'objets mathématiques et redondance graphique pour la lecture automatique de documents mathématiques

J.-Y. Toumit, S. Garcia-Salicetti, H. Emptoz
RFV – Insa de Lyon
bât. 403, 20, av. A. Einstein
69621 VILLEURBANNE CEDEX
Tel : 33.4.72.43.80.96, Fax : 33.4.72.43.80.97
e-mail : [jtoumit – sonia – emptoz] @rfv.insa-lyon.fr

Résumé

La rétroconversion des manuels scolaires est aujourd'hui un problème important pour les éditeurs ; nous travaillons actuellement dans ce contexte à la rétroconversion des documents et ouvrages de mathématiques. A notre connaissance, il n'y a que peu de travaux sur l'ensemble du document mathématique, seules des études concernant l'analyse des formules de mathématiques ont été développées à ce jour. C'est la raison pour laquelle nous posons le problème de la lecture automatique de ces documents.

Ceux-ci contiennent deux types d'informations de natures différentes : le texte et les objets mathématiques. Afin de traiter le texte plus efficacement, nous sommes conduits à séparer ces deux types d'informations ; dans cet article, nous nous intéressons particulièrement à cette étape qui peut être abordée comme un problème de segmentation multi-langages. Les méthodes classiques de segmentation ne donnant pas des résultats satisfaisants, nous avons été conduits à préconiser de nouvelles voies de segmentation physique et logique, de bas niveau.

Elles s'appuient en particulier sur la redondance graphique de caractères dans le texte, et la propagation autour de marqueurs que nous introduisons. Pour cela, une définition du texte mathématique est proposée, ainsi qu'une première classification des objets mathématiques le composant. Nous détaillons plus spécifiquement les techniques de redondance et la détection d'une certaine classe de formules mathématiques.

Ce travail est réalisé dans le cadre d'un contrat industriel avec la société PRITEC de Toulouse et un appui de l'ANVAR.

Mots-clefs

Rétroconversion, formules mathématiques, objets mathématiques, redondance graphique, segmentation

I. La lecture des documents mathématiques : un défi à relever

A. La problématique générale

La lecture automatique de documents mathématiques est à ce jour un problème très original, dans la mesure où les recherches effectuées s'attaquent plutôt au problème de la lecture automatique de formules mathématiques isolées ([1], [2]). Ainsi, le document mathématique est rarement vu dans son intégralité ([3], [4]). Aborder le document mathématique en tant que tel pose en effet plusieurs problèmes complexes, à un niveau plus global que celui de la reconnaissance de formules mathématiques: comment séparer le texte proprement mathématique du texte standard? Et, de surcroît, en amont de cette dernière question, comment définir le texte mathématique?

Ce sont les deux premières tâches que nous nous sommes données dans cet article.

B. Définitions

Notre but premier consiste à segmenter le document mathématique en texte standard et texte mathématique ; à ces fins, nous définissons le texte mathématique par rapport à la notion générique d'**objet** mathématique. L'**objet** mathématique est défini comme étant l'unité fondamentale du texte mathématique ; il a de nombreuses déclinaisons, des plus simples aux plus complexes ; ces dernières sont à leur tour composées d'**objets** agencés selon des règles propres à une certaine grammaire.

Dans la suite, nous appelons **texte mathématique** les objets mathématiques parsemés dans le document. Le texte mathématique comprend en effet des **formules**, des **abréviations** mathématiques (*sin* désigne ainsi la fonction sinus), des **signes** mathématiques particuliers (+, -, >, <, signe d'intégration, traits de