Методы семантической обработки математических документов

Герасимов Алексей Николаевич

Аспирант

Институт математики и механики имени Н. И. Лобачевского Казанского федерального университета, Казань, Россия

E-mail: sav241@mail.ru

Предложены методы семантического структурирования (например, [1]) математических статей, записанных в TEX-нотации. В частности, разработаны алгоритмы выделения библиографических описаний в списках литературы этих статей в нескольких вариантах стилевого оформления соответствующих файлов.

Наиболее сложными для обработки являются архивы научных статей. Структура таких документов в предложенных алгоритмах определялась на основе анализа шрифтового выделения и других элементов форматирования. Это позволило автоматизировать процесс выделения метаданных, в частности, библиографических данных.

Алгоритм извлечения элементов библиографических описаний состоит из: выделения блока библиографии; разделения его на отдельные библиографические записи (их признаками служат принятые правила оформления списка литературы, например, особенности нумерации); извлечения названия статьи, списка авторов и других выходных данных. Метод реализован на языке C#, извлечение элементов библиографических записей осуществляется с помощью паттернов регулярных выражений (например, [2]), специально разработанных для каждого журнала. В качестве тестового массива использовалась коллекция статей "Трудов Математического центра им. Н.И. Лобачевского" (более 40 томов).

Работа выполнена при финансовой поддержке РФФИ (проект 12-07-97018-р_поволжье) и РГНФ (проект 14-03-12004).

Литература

1. Елизаров А. М., Липачев Е. К., Малахальцев М. А. Веб-технологии для математика: Основы MathML. Практическое руководство. М.: Физматлит, 2010. 192 с.

2. Гойвертс Я., Левитан С. Регулярные выражения. Сборник

рецептов. СПб.: Символ-Плюс, 2010. 608 с.