УДК 004.93'1

Распознавание математических формул в PDF документе

И. А. Меньшиков

Московский физико-технический институт (государственный университет)

В работе рассматриваются вопросы распознавания математических формул в электронных документах научных статей PDF формата, их извлечения и сохранения в структурной форме в некоторой «базе знаний» математических формул, а так же нюансы соответствующей программной реализации.

Распознавание формул дает возможности: поиска научных документов по математическим формулам и подформулам, формирования базы знаний формул, переноса формул в другие документы, извлечения семантической информации из формулы, интегрирования с системами компьютерной алгебры.

В настоящее время не существует полномасштабного проекта анализа математических формул, бесплатного для использования и с открытым исходным кодом. Развитые проекты наподобие *Infty* [7] являются проприетарными, как и инструменты оптического распознавания, используемые подобными системами. Использование контента документа призвано снизить зависимость от затратной и неточной процедуры оптического распознавания текста.

Распознать формулу в документе значит задать границу области, содержащей все элементы формулы, распознать символы, входящие в данную область, сгруппировать символы в смысловые единицы, обозначенные геометрическим положением, - токены, идентифицировать связи между токенами, восстановить иерархическую структуру формулы. Полученная структура может быть транслирована в любой формат записи формулы.

Математические выражения представляют собой двумерную иерархическую структуру данных. Формулы могут содержать символы латинского алфавита, греческого алфавита специальные символы и графические объекты. Отметим, что изображение формул не формализовано строго, у представления формул отсутствует каноническая форма записи. Точное определение математического выражения требует задания формального списка математических символов в базе знаний формул.

Из PDF документа извлекаются глифы - объекты изображения или наборы графических команд, применяемые для визуализации символов. Особенность PDF документа заключается в том, что глифами могут выступать как элементы множества шрифтов, так и объекты изображений. Между глифами и символами не существует однозначного соответствия: символ может быть представлен

одним или несколькими глифами. Значения простых глифов определяются по таблице символов или с помощью геометрической интерпретации графических объектов. Композитные глифы определяются эвристическими правилами оптическим распознаванием. Задача актуализации таблиц соответствия глифов символам возложена на базе знаний формул.

По содержимому PDF документа нельзя непосредственно указать расположение формул на странице. В целом, PDF формат утрачивает структуру исходного файла при компиляции, отдельные графические элементы позиционируются абсолютными координатами на странице. Задача определения областей документа с формулами допускает различные интерпретации: геометрические, структурные, семантические. Чаще всего используется геометрическая сегментация формул на начальной стадии анализа для декомпозиции элементов страницы, например, метод разбиения профиля проекциями [6] *(projection profile cutting)*. Однако геометрическая сегментация часто не является достаточной: требуется дополнительный анализ для уточнения сегментации.

Сложность определения токенов заключается в использовании формульных переменных. Связи между токенами задаются операторами. В математической нотации существуют неявные операторы, заданные лишь относительным расположением частей формулы, распознавание которых осложняется отсутствием стандарта изображения неявных операторов.

Однако существуют свойства математической нотации, которые помогают идентифицировать и распознавать формулы [3]. Это геометрические свойства (отступы, выравнивание), свойства начертания (использование математических шрифтов, курсивное или жирное начертание), контекст. Эти свойства позволяют использовать эвристические правила анализа. К распознаванию формул адаптируются статистические методы распознавания: по обучающим экземплярам формул [2], гистограмме определенных параметров (например, пробельных расстояний) [1], специфические свойства могут использоваться в качестве векторов в методе опорных векторов [4], применяемом в алгоритме обучения. Используются методы геометрической декомпозиции формулы [8]. Методы могут рассматриваться в комбинациях друг с другом [4]. Для распознавания отношений между токенами используются разнообразные расширения грамматик (координатная грамматика Андерсона [1], двумерные грамматики, графовые грамматики).

В основе проекта используется *Java*-библиотека *Itext* [8], позволяющая извлечь всё содержимое PDF документа, кроме файлов шрифтов и изображений. Для извлечения бинарных объектов используется свободная библиотека *pdf.py* [9], написанная на языке *Python*. Далее, в проекте потребовался генератор метрических данных шрифтов [5], которые нужны для определения позиции символов на странице. Необходимы средства были найдены в библиотеках LaTeX компиляторов (например, *ttf2afm*).

В рамках данной работы проанализированы теоретические работы по распознаванию формул, которые проходят тестирование и улучшение на практике.

Литература

1. *Josef B. Baker et al.* A Linear Grammar Approach to Mathematical Formula Recognition from PDF. - University of Birmingham, 2009
2. *Jianming Jin et al.* Mathematical Formula Extraction. - Tianjin, China, 2001
3. *Xuan Hu et al.* Mathematical Formula Identification in PDF Documents. - 2011 International Conference on Document Analysis and Recognition.
4. *Fan Liu et al.* Mathematical Formula Detection in Heterogeneous Document Images
5. *С. В. Морогин.* Извлечение Математической Нотации из Документов Формата PDF.
6. *M. Alkalai, V. Sorge.* Issues in Mathematical Table Recognition. - University of Birmingham, 2012
7. http://www.inftyproject.org/en/index.html
8. http://itextpdf.com
9. https://github.com/mstamy2/PyPDF2/blob/master/PyPDF2/pdf.py