Mathematical Formula Recognition and Retrieval in PDF Documents

Xiaoyan Lin Advisor: Zhi Tang

Institute of Computer Science and Technology, Peking University, Beijing, China

Introduction

Objective

My research aims at developing effective approaches to recognize mathematical formulas in PDF documents and facilitate mathematical formulas retrieval.

Motivation

1. Mathematical formulas are difficult to extract from PDF documents

Most of the mathematical formulas in PDF documents are represented as images or complex objects, making mathematical formulas difficult to extract, manipulate and retrieve.

2. Information in PDF documents is richer than document images

Compared with document images, character and layout information in PDF files is richer and can be obtained directly through parsing the PDF files. In this sense, better results can be expected from formula recognition directly from PDF documents.

3. Mathematical formulas are difficult to retrieve

Mathematical formulas are objects with complex structures and distinct symbols. Thus, users can't search for relevant information centering on mathematical formulas in text-based search engines. Obstacles still exist in this area, e.g., building user-friendly interface for querying formulas, approximate matching between formulas, and sub-structure matching between formulas, etc.

Challenges

1. PDF documents are generated by different tools. It is challenging to build a parser for recognizing mathematical symbols from various versions of PDF documents.
2. The embedded formula is more difficult to identify than isolated formulas, because the embedded formulas are generally short expressions, which are difficult to discriminate from ordinary text.
3. Current structures analysis algorithms are mostly focused on the layout structure analysis of mathematical formula, whereas effective semantic analysis (or logical structure analysis) algorithms are few.

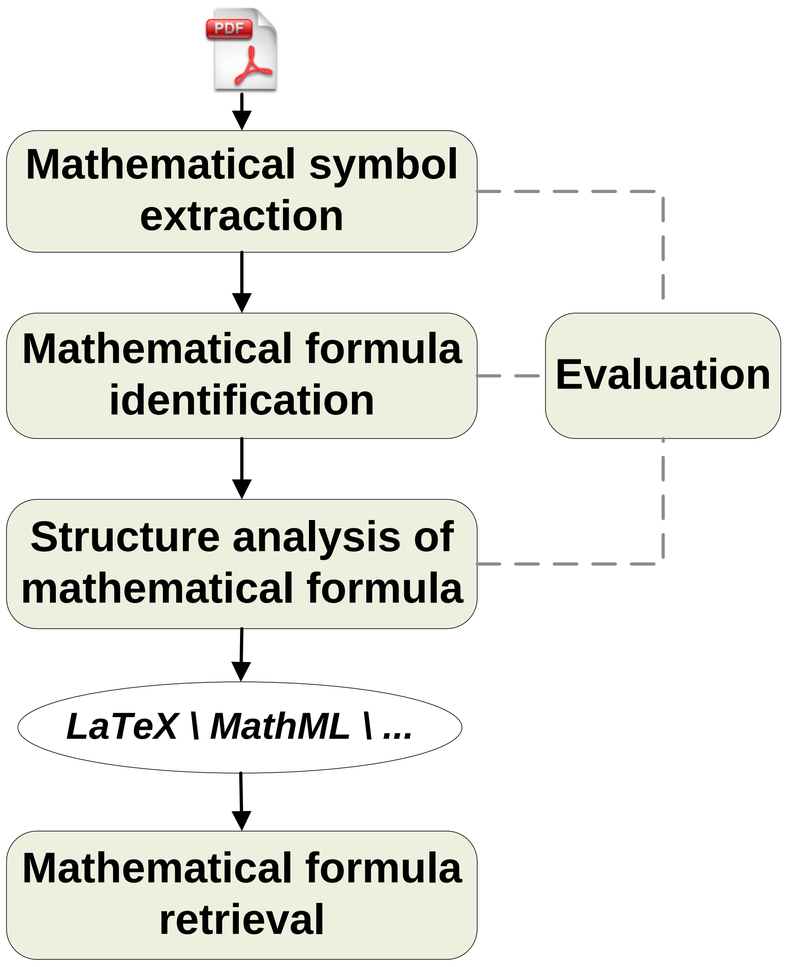

Figure 1. Overview of the proposed research

Proposed Research

Mathematical symbol extraction

Parse the text/graph/image objects from PDF documents, and then match them to mathematical symbols [1-2].

Mathematical formula identification

To improve the performance of formula identification, we will try to utilize the precise information of PDF and apply the Support Vector Machine (SVM) techniques to identify mathematical formulas in PDF documents.

Structure analysis

Adopt the existing formula structure analysis approaches for image documents with alteration in PDF documents [3]. Overcome the problems in traditional methods through fully utilizing the rich and accurate information in PDF files.

Mathematical formula retrieval

Design and implement a mathematical formula retrieval system which will be able to query by mathematical formulas, support approximate matching and sub-structure matching between formulas [4-5].

Evaluation

Build a ground-truth PDF documents dataset, for better comparing performances between different algorithms.

Progress

Mathematical formula identification

A hybrid method by combining rule-based and SVM-based methods was proposed to detect isolated mathematical expressions in PDF documents. Rule-based method is applied to extract embedded mathematical formulas.

Table 1. Result of the isolated formula identification

Method	Precision	Recall	F1
Rule-based	90.54%	90.66%	90.60%
Learning-based	94.33%	97.01%	95.64%
Rule-based + Learning-based	**94.45%**	**97.91%**	**96.14%**

Table 2. Result of the embedded formula identification

Method	Precision	Recall	F1
Rule-based	83.05%	84.18%	83.61%

Structure analysis of mathematical formula

The baseline structure analysis algorithm is adopted to analyze mathematical formulas structure in PDF documents. The parse trees representing the layout structures of the math symbols are created and exported into MathML (in Presentation Markup).

Bibliography

[1] J. Baker, A. Sexton, V. Sorge. "A linear grammar approach to mathematical formula fecognition from PDF", In Proc. of Mathematical Knowledge Management 2009.

[2] X.Y. Lin, L.C. Gao, Z. Tang, X.F. Lin and X. Hu. "Mathematical formula identification in PDF documents", To appear in ICDAR 2011.

[3] J. Baker, A. Sexton, and V. Sorge. "Faithful mathematical formula recognition from PDF documents", In Proc. of IAPR International Workshop on Document Analysis Systems 2010.

[4] R. Miner and R. Munavalli, "An approach to mathematical search through query formulation and data normalization", In Towards Mechanized Mathematical Assistants, MKM 2007.

[5] R. Zanibbi and B. Yuan. "Keyword and image-based retrieval for mathematical expressions", Document Recognition and Retrieval 2011.