Annotation of mathematical formulas in PDF documents

Konstantin Nikolaev
Federal State Institution of the Federal Research Center NIISI RAS
Kazan, Russia
konnikolaeff@yandex.ru

Olga Nevzorova
Kazan Federal University
Kazan, Russia
onevzoro@gmail.com

***Abstract*—This article provides an overview of existing solutions for semantic analysis of mathematical documents, and also presents a method for automatic semantic analysis of documents in PDF format. This method searches for local variables in the text of the article, extracts their definitions and connects concepts with formulas. The advantage of the method over the existing ones is independence from the markup of the original PDF document, which expands the scope of the method. We provide estimates of recall, precision and F-measure for algorithms for finding variables and linking local variables with formulas. The resulting semantic markup of the document will be used to create a collection of documents suitable for the semantic formula search service, which is part of the set of services of the Lobachevskii-DML digital publishing system.**

Keywords—semantic analysis, PDF, document processing, scientific journals, Lobachevskii-DML

I. Introduction

Semantic search is focused on searching in collections of semantic publications, which are documents with semantic markup of text components. Mathematical texts are highly structured, with the presence of fixed semantics components, such as theorems, proofs, formulas, etc.

The task of searching for documents on mathematical formulas is relevant for conducting scientific research, preparing articles, studying mathematical disciplines. The paper [1] describes a semantic search engine for mathematical formulas, which uses a data set based on a collection of scientific articles of the journal "Izvestiya Vuzov. Mathematics" for 1997-2009. This article discusses new improved algorithms for constructing a data set for semantic search by mathematical formulas, which will qualitatively improve the search results.

Mathematical search by formulas can be divided into two categories – search for formulas by structure and by content. Searching for formulas by structure comes down to getting a list of formulas that partially or completely match the structure of the formula specified in the search query. This approach does not take into account the semantics of formulas.

More effective, but difficult to implement, is the search for mathematical articles on the content of the formula. To determine the content of the formula, it is necessary to identify the variables of that formula, and to define mathematical concepts denoted by variables. Additional difficulties are caused by a variety of templates for the design of mathematical documents for different information systems of scientific journals. Currently, the most popular formats for presenting mathematical formulas in scientific articles are: graphic image (articles in pdf format); formulas in Microsoft Word editor; LaTeX format; MathML format. Collections of articles in digital mathematical libraries are presented mainly in PDF format. Text recognition, and, in particular, mathematical expressions, is the main task of this study. Mathematical formulas extracted during recognition in scientific articles and their descriptions are the source data for building an improved data set for a mathematical search engine.

II. Method Of Semantic Annotation Of Formulas In A PDF Document

Most of the existing solutions of semantic annotation of documents strongly depend on the input document, its format and structure [2-6]. The article proposes a universal method for determining the structure of a PDF document and linking variables in the text with the main formulas for determining the semantic content of the document. The data set built on the basis of the developed algorithms will be used to improve the quality of semantic search for mathematical formulas.

Semantic annotation of a formula consists in extracting a formula that meets special requirements from the text of a mathematical article, followed by an analysis of its structural elements and linking the selected formula variables with the legends given in the textual context of the formula.

The main task of semantic annotation of formulas is to develop a software solution allowing to identify a set of variables in the formulas of a mathematical document, and associate variables with mathematical concepts using mathematical ontology. The resulting semantic markup of the document will make possible the creation of a collection of documents suitable for the semantic formula search service, which is part of the set of services of the Lobachevskii-DML digital library.

To solve the problem of semantic annotation of formulas in PDF documents, the following tasks were solved:

- Splitting the document into blocks.
- Highlighting the main formulas and text blocks.
- Search for variables in text blocks.
- Recognition of the main formulas and local variables.
- Linking formulas and local variables.
- Markup of mathematical concepts in text blocks based on OntoMathPRO ontology.
- Linking the selected concepts with the variables of the formula.

These tasks were performed using python programming language. The division into blocks was carried out by analyzing the markup of the document. The main formulas are the formulas highlighted in a separate paragraph. Local variables are located in text blocks. Linking local variables and main formulas was performed by searching for common formulas. Fig. 1 shows an example of linking local variable to the main formula.

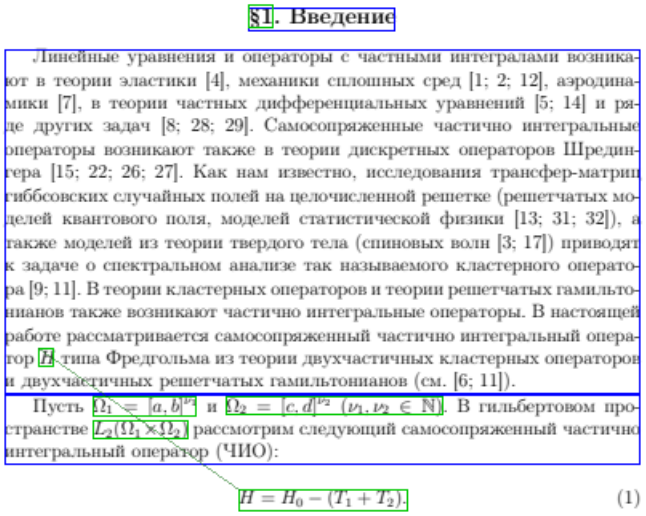

Fig. 1. Linking the main formula and the variable

Recognition of concepts in the text was performed using the OntoMathPRO ontology concept extraction algorithm. This method is used in the preparation of mathematical educational courses at Kazan Federal University. The main idea of this algorithm is to search for all chains of words in a sentence and compare them with similarly constructed chains in concepts of ontology. The algorithm accepts documents in html format as input, therefore, a method for generating an intermediate html representation of a PDF document was created for its application in this task. In the tags of such a representation, a text representation of the source document was obtained, indicating the number of the block. Fig. 2 shows an example of recognizing concepts from the OntoMathPRO anthology in a text block of a document.

Fig. 2. Recognized concepts in the text of the document

The developed algorithm for linking local variables and main formulas has the following estimates: precision - 0.81, recall – 0,72 , F-measure – 0,75.

Table 1 shows examples of the main formulas and related local variables.

TABLE I. EXAMPLES OF RECOGNIZED FORMULAS AND VARIABLES

Main formula	***Variable***	***Ontology concept***
$\int_{\Omega_j} \varphi_j(\xi) d\mu_j(\xi) = 0, \quad \int_{\Omega_j} \varphi_j^2(\xi) d\mu_j(\xi) = 1$	$\mu_j(\cdot)$	Lebesgue measure
$t \mapsto \frac{n!}{\sqrt{\sum_{i=1}^{n}(x_i - x_0)^2}} \cdot \mathrm{mes}\,(U(t)),$	$\mathrm{mes}\,(U(t))$	Lebesgue measure
$H = H_0 - (T_1 + T_2).$	H	Integral operator

III. CONCLUSION

The article presents a method of semantic annotation of mathematical documents in PDF format. A method for determining the structure of a document by dividing it into blocks with text and main formulas is described. A method of linking local variables in the text with the main forms has been developed. With the help of the method of marking mathematical concepts in the text, a semantic representation of the main formulas is formed. The developed method is used to prepare a data set for a semantic search engine using formulas in the Lobachevskii-DML digital library.

Future developments are related to the expansion of the method's capabilities, in particular, for linking many variables of a mathematical function, as well as the development of a method for filtering concepts found in a sentence to increase the accuracy of the annotation method. Another direction for increasing the universality of the method can be considered planning the implementation of an OCR module for text recognition in scanned PDF, which is especially relevant for mathematical articles published in the pre-digital era.

ACKNOWLEDGMENT

The study was carried out with the support of the Russian Science Foundation, project No. 21-11-00105.

REFERENCES

[1] Nevzorova O. The semantic context models of mathematical formulas in scientific papers / O. Nevzorova, A. Kirillovich, V. Nevzorov, K. Nikolaev // CEUR Workshop Proceedings. – 2018. – Vol. 2277. – P. 33-40.

[2] Bertin M. Hybrid Approach for the Semantic Processing of Scientific Papers / M. Bertin, I. Atanassova // Semantic Publishing Challenge Track in 11 th European Semantic Web Conference (ESWC 2014). – 2014.

[3] Ciancarini P. Semantic annotation of scholarly documents and citations / P. Ciancarini, A. Di Iorio, A. G. Nuzzolese et al. // Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). – 2013. – Vol. 8249 LNAI. – P. 336-347.

[4] Ronzano F. Semantify CEUR-WS proceedings: Towards the automatic generation of highly descriptive scholarly publishing linked datasets / F. Ronzano, G. C. Del Bosque, H. Saggion // Communications in Computer and Information Science. – 2014. – Vol. 475. – P. 83-88.

[5] Ahmad R. Information extraction from PDF sources based on rule-based system using integrated formats / R. Ahmad, M. T. Afzal, M. A. Qadir // Communications in Computer and Information Science. – 2016. – Vol. 641. – P. 293-308.

[6] Greiner-Petter A. Math-word embedding in math search and semantic extraction / A. Greiner-Petter, A. Youssef, T. Ruas [et al.] // Scientometrics. – 2020. – Vol. 125. (3). – P. 3017-3046.