Web-based visualization of semantic annotation of mathematical PDF documents

Konstantin Nikolaev[1][0000-0003-3204-238X]

[1] Federal State Institution of the Federal Research Center NIISI RAS
konnikolaeff@yandex.ru

Abstract. This paper presents a method for visualizing semantic markup of texts and formulas of mathematical PDF documents. To form the markup visualization, the document structure is analyzed, formulas and in-text variables are searched, semantic markup of text blocks with placement with the formation of links to the description of concepts from the ontology of professional mathematics OntoMathPRO [1] is carried out. Visualization of markup results is performed using web technologies (CSS, JavaScript) and is available for viewing from any device. The visualization method will be implemented in the OntoMath ecosystem as a component of the Lobachevskii-DML publishing system. We also plan to apply the results of the method in the semantic search service for formulas and concepts in PDF documents.

Keywords: PDF document, semantic markup, formula search, web technologies.

1 Introduction

Semantic document markup is a highly demanded task in the scientific community. Firstly, documents enriched with semantic markup enhance the process of knowledge assimilation for readers. Secondly, semantically enriched documents on the Internet allow search engines to better analyze their content and meaning, increasing the likelihood of these documents appearing in search results. Furthermore, semantic markup provides the foundation for implementing specialized search systems, such as semantic formula search.

For the successful application of the advantages of documents subjected to the semantic markup process, as well as for debugging the markup method itself, it is necessary to develop methods for visualizing the markup results in a format convenient for the end user. The most common way of transmitting information to the user nowadays are web technologies. The advantages of web technologies include ease of development, versatility (web services are available from most modern mobile and personal computers) and undemanding user resources (all calculations are usually performed on the server side).

In addition, an important requirement for the means of visualizing the results of semantic markup is the ease of implementation in other web services, such as publishing systems and electronic scientific libraries.

Related work.

There are a number of works aimed at visualizing certain semantic data (ontologies, semantic graphs). One of the areas of knowledge in which data visualization tasks are particularly relevant is biology. For example, in [2], the authors offer a service for forming a graphical representation of ontological relationships between different plant species. To do this, the ontology is transformed into a graph data model with subsequent visualization using the library D3.js. In [3], the MetDraw service is presented, which forms graphical representations of complex metabolic models. In [4], the PlantMedSuite service is presented, which forms graphical representations for plant metabolomics, and allows for the analysis of such systems in an interactive mode.

In addition, a number of works are aimed at visualizing historical and geographical data. The SEMAP project [5] provides interactive access to a cloud of data collected from many Spanish museums. To do this, a combination of data from Google Maps and the Geonames API is used. In [6], a tool is proposed for visualizing heterogeneous data with the possibility of clarifying the desired data by time intervals and geographical location. Records from 25 datasets are used as input data. In [7], an interesting approach is proposed to display large amounts of data in the form of blocks, the sizes of which depend on the number of records in this block, with the possibility of revealing blocks of interest to the user. The authors [8] propose the GeoTemCo service, which uses Dynamic Delaunay Triangulations to cluster densely located elements on the map into so-called circle groups, for convenient management of the visualization process of geographical data.

Most of the solutions related to the visualization of semantic data inside documents are limited to enriching the text with hyperlinks to more detailed information about related concepts. This paper proposes a method for visualizing the results of semantic markup in graphical mode, with the possibility of analyzing the relationships between the mathematical components of the document.

2 Automatic Semantic Annotation Method for PDF Documents

In the study [9] we introduce an algorithm for the automatic semantic annotation of mathematical PDF documents. The result of this method is the semantic enrichment of formulas in the document. Semantic enrichment refers to a set of concepts that describe the components of a formula, such as functions and variables within the text blocks of the document. The document segmentation into text blocks and main formulas is performed through the analysis of the document markup. Fig. 1 illustrates an example of the document segmentation into text blocks and main formulas. The semantic annotation of text blocks is carried out using a modified method for annotating HTML documents. The original version of this method is used in the preparation of educational

materials for a distance learning course in school geometry at Kazan Federal University [10]. Fig. 2 presents the algorithm for semantic annotation of PDF documents. The main stages of the algorithm can be divided into document preprocessing (identifying and numbering text blocks, text normalization), concepts preprocessing from the OntoMathPro ontology [1], and concept extraction from the text using the Jaccard measure.

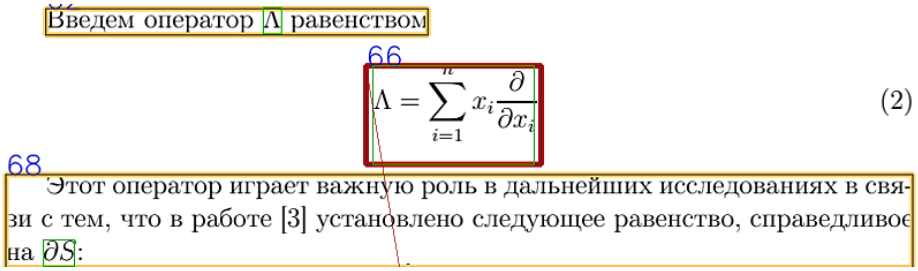


Fig. 1. Segmentation of the document into text blocks and main formulas

Fig. 3 shows an example of semantic annotation of a PDF document. The words in the document text that correspond to concepts from the OntoMathPro ontology are highlighted in color. The corresponding ontology concepts are indicated in parentheses (for example, the word “triangle” is associated with the concept “Triangle”).

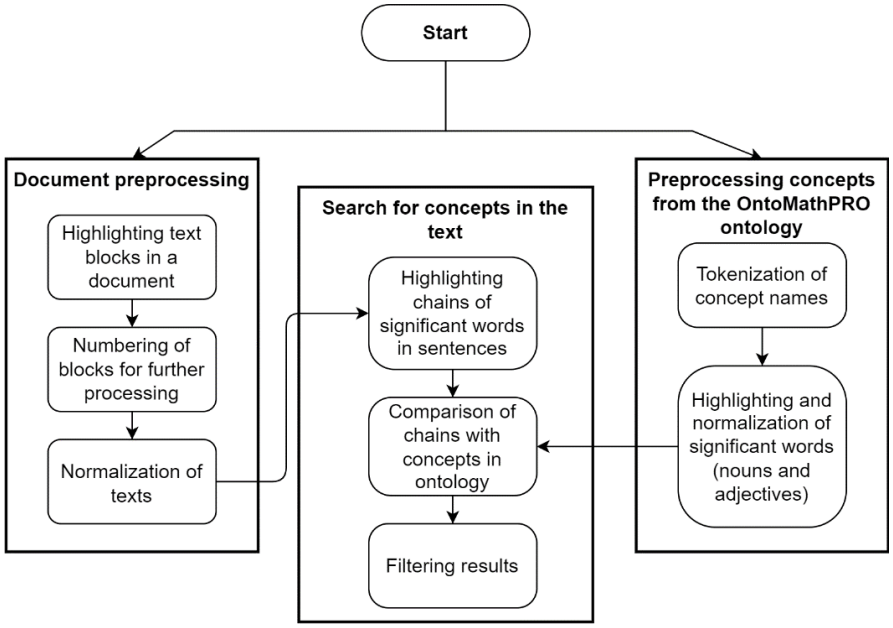
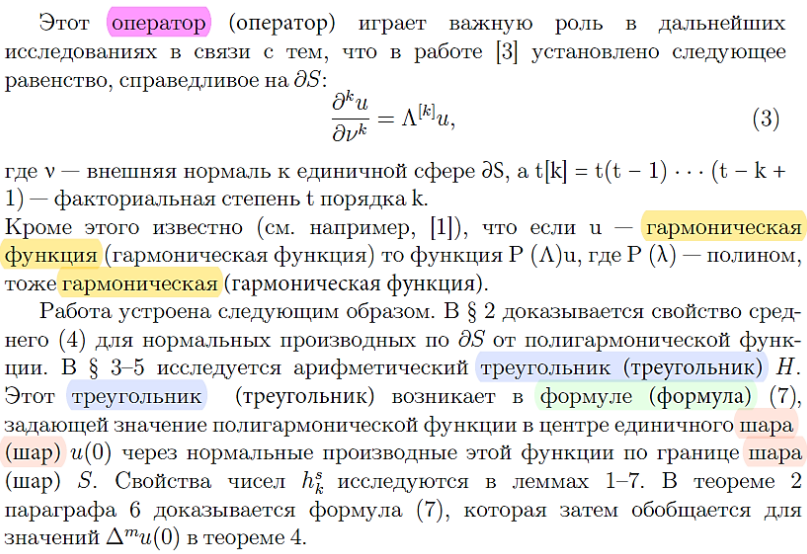


Fig. 2. The algorithm for concept annotation in PDF document texts



Этот оператор (оператор) играет важную роль в дальнейших исследованиях в связи с тем, что в работе [3] установлено следующее равенство, справедливое на ∂S:

$$\frac{\partial^k u}{\partial \nu^k} = \Lambda^{[k]} u, \tag{3}$$

где ν — внешняя нормаль к единичной сфере ∂S, а t[k] = t(t − 1) · · · (t − k + 1) — факториальная степень t порядка k.
Кроме этого известно (см. например, [1]), что если u — гармоническая функция (гармоническая функция) то функция P (Λ)u, где P (λ) — полином, тоже гармоническая (гармоническая функция).

Работа устроена следующим образом. В § 2 доказывается свойство среднего (4) для нормальных производных по ∂S от полигармонической функции. В § 3–5 исследуется арифметический треугольник (треугольник) H. Этот треугольник (треугольник) возникает в формуле (формула) (7), задающей значение полигармонической функции в центре единичного шара (шар) $u(0)$ через нормальные производные этой функции по границе шара (шар) S. Свойства чисел h_k^s исследуются в леммах 1–7. В теореме 2 параграфа 6 доказывается формула (7), которая затем обобщается для значений $\Delta^m u(0)$ в теореме 4.

Fig. 3. An example of semantic annotation of a document. Words in the text that correspond to concepts from the OntoMathPro ontology are highlighted in color (the corresponding concept is indicated in parentheses)

The output of the method for semantic annotation of PDF documents is a set of terms linked to words in text blocks, as well as a separate set of terms for each main formula.

Since the results of this method will be embedded in the Lobachevskii-DML publishing system, it is necessary to develop a method for visualizing the results of the method using web technologies: HTML, CSS, JavaScript. In this article, we present a method for visualizing terms and LaTeX formulas, as well as the results of semantic annotation of text blocks and formulas.

3 Visualization of Annotation Results

This method is being developed for integration into OntoMath ecosystem, which represents a set of data sources and services for organizing the study and application of professional mathematics concepts, as well as for the semantic enrichment of documents in the Lobachevskii-DML publishing system.

The operation of the visualization method for the results of semantic annotation consists of three stages: displaying recognized LaTeX representations of local variables in the text and main formulas, displaying recognized concepts in text blocks, and displaying the relationships between local variables and main formulas. To visualize this data, an HTML document is generated, which is based on images of the pages of the original document.

3.1 Displaying Recognized Concepts in the Document Text

Text blocks (representing recognized concepts) and inline formulas in LaTeX format, visualized using the MathJax library [11], are placed over the document pages. The positioning and sizes of these blocks are determined during the process of PDF document markup. Fig. 4 and 5 provide examples of displaying recognized concepts in the documents. Currently, the recognized words are highlighted with dashed rectangles, and the corresponding concept is hidden by default. When hovering the mouse cursor over a highlighted word, the corresponding concept is displayed. The web page with the document also includes a toggle switch to control the display of recognized concepts. Clicking on a concept navigates to the concept page on the resource [12]. Adding links to detailed information about the concept allows the user to access more detailed information about the concept within the text.

Введем оператор Λ равенством

$$\Lambda = \sum_{i=1}^{n} x_i \frac{\partial}{\partial x_i}. \tag{2}$$

Этот оператор играет важную роль в дальнейших исследованиях в связи с тем, что в работе [3] установлено следующее равенство, справедливое на ∂S:

$$\frac{\partial^k u}{\partial \nu^k} = \Lambda^{[k]} u, \tag{3}$$

где ν — внешняя нормаль к единичной сфере ∂S, а $t^{[k]} = t(t-1)\cdots(t-k+1)$ — факториальная степень t порядка k.

Кроме этого известно (см. например, [1]), что если u — гармоническая функция, то функция $P(\Lambda)u$, где $P(\lambda)$ — полином, тоже гармоническая.

Работа устроена следующим образом. В § 2 доказывается свойство среднего (4) для нормальных производных по ∂S от полигармонической функции. В §3–5 исследуется арифметический треугольник H. Этот треугольник возникает в формуле (7), задающей значение полигармонической функции в центре единичного шара $u(0)$ через нормальные производные этой функции по границе шара S. Свойства чисел h_k^s исследуются в леммах 1–7. В теореме 2 параграфа 6 доказывается формула (7), которая затем обобщается для значений $\Delta^m u(0)$ в теореме 4.

Fig. 4. Example of displaying recognized concepts in documents: positions of recognized words.

Введем оператор Λ равенство

$$\Lambda = \sum_{i=1}^{n} x_i \frac{\partial}{\partial x_i}. \tag{2}$$

Этот оператор играет важную роль в дальнейших исследованиях в связи с тем, что в работе [3] установлено следующее равенство справедливое на ∂S:

$$\frac{\partial^k u}{\partial \nu^k} = \Lambda^{[k]} u, \tag{3}$$

где ν — внешняя нормаль к единичной сфере ∂S, а $t^{[k]} = t(t-1)\cdots(t-k+1)$ — факториальная степень t порядка k.

Кроме этого известно (см. например, [1]), что если u — гармоническая функция, то функция $P(\Lambda)u$, где $P(\lambda)$ — полином, тоже гармоническая.

Работа устроена следующим образом. В § 2 доказывается свойство среднего (4) для нормальных производных по ∂S от полигармонической функции. В §3–5 исследуется арифметический треугольник H. Этот треугольник возникает в формула (7), задающей значение полигармонической функции в центре единичного шара $u(0)$ через нормальные производные этой функции по границе шара S. Свойства число h_k^s исследуются в леммах 1–7. В теорема 2 параграфа 6 доказывается формула (7), которая затем обобщается для Значение $\Delta^m u(0)$ в теореме 4.

Fig. 5. Example of displaying recognized concepts in documents: ontology concepts corresponding to highlighted words.

3.2 Displaying Recognized Variable Formulas in LaTeX Format

To recognize LaTeX formulas in situations where the symbols in the formula are not recognized by the library for processing PDF documents (pdfminer), the pic2tex library is used. This library utilizes a pre-trained neural network to recognize LaTeX formulas in images. To improve the quality of formula recognition, this library is applied to each formula and variable in the text, and the most accurate option is chosen between the output of pic2tex and the set of symbols from the original document. As a result, a set of LaTeX formulas and their positions in the document are generated. The boundaries of the formula and variable are highlighted with dashed rectangles, and the corresponding LaTeX formulas are hidden by default. When hovering over a formula or variable in the text, the corresponding LaTeX formula is displayed. Fig. 6 and 7 provide an example of displaying recognized LaTeX representations in main formulas and local variables.

Введем оператор Λ равенством

$$\Lambda = \sum_{i=1}^{n} x_i \frac{\partial}{\partial x_i}. \tag{2}$$

Этот оператор играет важную роль в дальнейших исследованиях в связи с тем, что в работе [3] установлено следующее равенство, справедливое на ∂S:

$$\frac{\partial^k u}{\partial \nu^k} = \Lambda^{[k]} u, \tag{3}$$

где ν — внешняя нормаль к единичной сфере ∂S, а $t^{[k]} = t(t-1)\cdots(t-k+1)$ — факториальная степень t порядка k.

Кроме этого известно (см. например, [1]), что если u — гармоническая функция, то функция $P(\Lambda)u$ где $P(\lambda)$ — полином, тоже гармоническая.

Работа устроена следующим образом. В § 2 доказывается свойство среднего (4) для нормальных производных по ∂S от полигармонической функции. В § 3–5 исследуется арифметический треугольник H. Этот треугольник возникает в формуле (7), задающей значение полигармонической функции в центре единичного шара $u(0)$ через нормальные производные этой функции по границе шара S. Свойства чисел h_k^s исследуются в леммах 1–7. В теореме 2 параграфа 6 доказывается формула (7), которая затем обобщается для значений $\Delta^m u(0)$ в теореме 4.

Fig. 6. An example of displaying recognized LaTeX representations in main formulas and local variables: position of recognized formulas and variables.

As can be seen from Fig. 6 and 7, some in the text have been recognized incorrectly. Incorrect recognition of LaTeX formulas hampers the process of linking local variables and main formulas, which can compromise the accuracy of determining the semantic content of the main formulas. The display of recognized LaTeX formulas is necessary for further debugging of the methods for formula and variable recognition and linking in the text.

3.3 Displaying Relations Between Local Variables and Main Formulas

To visualize the found relationships between local variables and the main formulas, the method of drawing lines using CSS tools is used. The <svg> tag and nested <line> tags for individual lines are used for rendering. The lines connect the main formulas and the local variables that are part of the main formula. If a corresponding concept is found for a local variable, the line color is set to green. Otherwise, the line color is red. Fig. 8 shows an example of displaying relationships between the main formula and local variables. To simplify perception, the recognized concepts are hidden.

Введем оператор Λ равенством

$$\Lambda = \sum_{i=1}^{n} x_i \frac{\partial}{\partial x_i}. \tag{2}$$

Этот оператор играет важную роль в дальнейших исследованиях в связи с тем, что в работе [3] установлено следующее равенство, справедливое на ∂S

$$\frac{\partial^k u}{\partial \nu^k} = \Lambda^{[k]} u, \tag{3}$$

где $\mathcal{L}$ — внешняя нормаль к единичной сфере ∂S, а $t^{[k]} = t(t-1)\cdots(t-k+1)$ — факториальная степень t порядка k.

Кроме этого известно (см. например, [1]), что если u — гармоническая функция, то функция $P(\Lambda)u$ где $P(\lambda)$ — полином, тоже гармоническая.

Работа устроена следующим образом. В §2 доказывается свойство среднего (4) для нормальных производных по ∇ от полигармонической функции. В §3–5 исследуется арифметический треугольник H. Этот треугольник возникает в формуле (7), задающей значение полигармонической функции в центре единичного шара $u(0)$ через нормальные производные этой функции по границе шара S. Свойства чисел h_k^s исследуются в леммах 1–7. В теореме 2 параграфа 6 доказывается формула (7), которая затем обобщается для значений $\Delta^m u(0)$ в теореме 4.

Fig. 7. An example of displaying recognized LaTeX representations in main formulas and local variables: recognized LaTeX formulas

Здесь действия операторов H_0, T_1 и T_2 определяются по формулам

$$H_0 f(x,y) = k_0(x,y) f(x,y), \quad f \in L_2(\Omega_1 \times \Omega_2),$$
$$T_1 f(x,y) = \int_{\Omega_1} (\gamma_0 + \gamma \varphi_1(x)\varphi_1(s))\, f(s,y) d\mu_1(s), \quad \gamma_0 \geq 0, \quad \gamma > 0,$$
$$T_2 f(x,y) = \int_{\Omega_2} (\mu_0 + \mu \varphi_2(y)\varphi_2(t))\, f(x,t) d\mu_2(t), \quad \mu_0 \geq 0, \quad \mu > 0,$$

где $k_0(x,y)$ — неотрицательная непрерывная функция на $\Omega_1 \times \Omega_2$, $\varphi_j(\cdot)$ — веществознозначная непрерывная функция на Ω_j,

$$\int_{\Omega_j} \varphi_j(\xi) d\mu_j(\xi) = 0, \quad \int_{\Omega_j} \varphi_j^2(\xi) d\mu_j(\xi) = 1$$

и $\mu_j(\cdot)$ — мера Лебега на Ω_j, $j = 1, 2$. Таким образом, оператор H из (1) зависит от четырех параметров γ_0, γ, μ_0 и μ, т. е. $H = H(\gamma_0, \gamma, \mu_0, \mu)$

Fig. 8. An example of displaying relationships between the main formula and local variables. The green lines show links to local variables associated with concepts in the text. The red lines show connections with local variables without recognized concepts.

Performing three stages of the markup results visualization method forms a web page that is easily embedded in other web services. Saving the rendered pages of the original PDF document into separate files simplifies the overlay of the components of the visualization method and eliminates dependence on the features of libraries for rendering PDF documents in the browser.

4 Conclusion

The method presented in this article automatically generates a representation of the results obtained from processing PDF documents, providing users of the markup method the ability to analyze recognized concepts and determine the semantics of main formulas. Implementing this method using web technologies allows easy integration into a web application for managing the semantic markup of mathematical PDF documents. The results of this method will be integrated into the Lobachevskii-DML publishing system as a database for the semantic search service based on formulas and concepts.

In the future, it is planned to add controls for hiding, displaying and customizing individual overlay elements, as well as expanding the dataset with mathematical concepts, for more complete semantic markup of the document and, as a result, improving the quality of linking local variables and main formulas.

Acknowledgement.

Funding: The article was prepared within the framework of the government task of the Federal State Institution of the Federal Research Center NIISI RAS for 2022-2024 (topic No. 0580-2022-0014 (FNEF-2022-0014)).

References

1. Elizarov, A.M., Kirillovich, A. V., Lipachev, E.K., Nevzorova, O.A.: OntoMathPRO: An Ontology of Mathematical Knowledge. Dokl. Math. 106, 429–435 (2022). https://doi.org/10.1134/S1064562422700016.
2. Mohamad-Matrol, A.A., Chang, S.-W., Abu, A.: Plant data visualisation using network graphs. PeerJ. 6, e5579 (2018). https://doi.org/10.7717/peerj.5579.
3. Jensen, P.A., Papin, J.A.: MetDraw: automated visualization of genome-scale metabolic network reconstructions and high-throughput data. Bioinformatics. 30, 1327–1328 (2014). https://doi.org/10.1093/bioinformatics/btt758.
4. Liu, Y., Liu, H.-Z., Chen, D.-K., Zeng, H.-Y., Chen, Y.-L., Yao, N.: PlantMetSuite: A User-Friendly Web-Based Tool for Metabolomics Analysis and Visualisation. Plants. 12, 2880 (2023). https://doi.org/10.3390/plants12152880.
5. Sevilla, J., Casanova, P., Samper, J.J., Portales, C.: SeMap: A project on semantisation and visualisation of cultural heritage data. ACM Int. Conf. Proceeding Ser. (2022). https://doi.org/10.1145/3544538.3544639.
6. Simon, R., Isaksen, L., Barker, E., Cañamares de Soto, P.: Peripleo: a Tool for Exploring Heterogeneous Data through the Dimensions of Space and Time. Code4Lib. 31, 1–8 (2016).
7. Xu, W., Esteva, M., Jain, S.D., Jain, V.: Interactive visualization for curatorial analysis of large digital collection. Inf. Vis. 13, 159–183 (2014). https://doi.org/10.1177/1473871612473590.
8. Jänicke, S., Heine, C., Scheuermann, G.: GeoTemCo: Comparative Visualization of Geospatial-Temporal Data with Clutter Removal Based on Dynamic Delaunay Triangulations. Commun. Comput. Inf. Sci. 359 CCIS, 160–175 (2013). https://doi.org/10.1007/978-3-642-38241-3_11.

9. Nevzorova, O.A., Nikolaev, K.S.: Semantic Annotation of Mathematical Formulas in PDF-Documents. Russ. Digit. Libr. J. 25, 616–639 (2023). https://doi.org/10.26907/1562-5419-2022-25-6-616-639.
10. Falileeva, M., Gajfullina, A., Sal'nikova, E., Safina, A., Nikolaev, K., Tarasov, T.: Tekhnologiya resheniya planimetricheskih zadach, https://edu.kpfu.ru/course/view.php?id=2652, last accessed 2023/06/13.
11. MathJax: Beautiful and accessible math in all browsers, https://www.mathjax.org/, last accessed 2023/06/13.
12. OntoMathpro Ontology classes, http://ontomathpro.org/ontology/, last accessed 2023/06/13.