Semantic Formula Search in Digital Mathematical Libraries

Alexander Elizarov, Alexander Kirillovich, Evgeny Lipachev
N.I. Lobachevskii Institute of Mathematics and Mechanics
Kazan (Volga Region) Federal University
Kazan, Russian Federation
amelizarov@gmail.com, alik.kirillovich@gmail.com, elipachev@gmail.com

Olga Nevzorova
Research Institute of Applied Semiotics
Tatarstan Academy of Sciences
Kazan (Volga Region) Federal University
Kazan, Russian Federation
onevzoro@gmail.com

***Abstract*— We are presenting semantic methods of search for mathematical objects in scientific publications. In particular, methods of search for mathematical formulas, as well as methods based on the logical structure of mathematical documents, are being discussed here. Based on the digital mathematical library Lobachevskii DML, created at Kazan Federal University in 2017, declared as Lobachevsky Year, we developed and tested new methods of search in digital collections of mathematical documents.**

Keywords*— *mathematical content search; semantic search; mathematical formulas search; information retrieval; World Digital Mathematical Library; OntoMath ecosystem

I. Introduction

Search in mathematical documents is today a topical and rapidly developing field of research (see, for example, [1, 2]). As is known, the meaning of mathematical texts is mostly determined by their formula content: authors of many works note that about 80% of the informative information of mathematical scientific articles are contained in the formulas presented in them. Information systems of general purpose, carrying out a search for scientific publications, only implement a search for the text content of articles. Examples of such systems are Google Scholar, Microsoft Academic Search, CiteseerX. For semantic processing of collections of mathematical documents through such search services it is obviously not enough. Therefore, it is necessary to develop specialized search systems for mathematical documents that support search by mathematical objects and formulas [1–3].

Solving the problem of mathematical formulas retrieval also has a great independent significance. In addition, the solution to this problem is the basis of constructing digital mathematical libraries (DML). Numerous attempts have been made to solve this problem, but none has found application and recognition in the broad mathematical community. At the same time, there is still no generally accepted agreement on the mathematical search format to be used by the library systems or Google Scholar [4]. In the process of integrating existing DML into larger projects (such as, for example, the European Digital Mathematics Library (EuDML)), the solution to the problem of mathematical formulae retrieval becomes more and more relevant: without the support of mathematical search, the efficiency of DML is significantly reduced [5].

Currently, only a few of the DML support search for mathematical objects. As noted in [1], exceptions are digital libraries EuDML and DML-CZ which use the Math Indexer and Searcher (MIaS) search tool which is developed at the Faculty of Informatics of the University of Masaryk. We also discuss the search services for collections of digital mathematical documents we created while developing the OntoMath digital ecosystem [6] and forming the digital mathematical library Lobachevskii DML.

The problem under discussion became even more urgent due to emergence and active development of the World Digital Mathematical Library (WDML) project [7, 8].

When developing methods of search for mathematical documents, it is necessary to take into account not only the semantics, but also the format of the document presentation (see, for example, [9]). The methods that currently exist are focused mainly on search for documents in .tex and docx-formats (see, for example, [2]). In this case, the search query itself should be expressed in the form of text. We note in this connection that new methods of multimodal search by formulas allow us to formulate the query in the graphic form [10].

Let us briefly describe the structure of the work. In section II we indicate the approaches closest to ours. In Section III we give an overview of the OntoMath digital ecosystem which consists of a set of ontologies on the basis of which a semantic search service is organized according to formulas. We present this service in section IV. The last section is devoted to the digital mathematical library Lobachevskii Digital Mathematics Library, within the framework of which our approaches are being realized.

II. Related Works

To date, a large amount of research has been done to find mathematical formulas in electronic libraries and the Internet: several mathematical search engines (MSE) have been created: MathDex, EgoMath, LATEXSearch, LeActiveMath and

MathWebSearch. The named systems are described in [1, 4]. Among the examples of information retrieval systems working in local mathematical collections are the search engine of the archive of the journal "Lobachevskii Journal of Mathematics (1998-2007)" [11], KWARC MathWebSearch [12], and MIaS (Math Indexer and Searcher) [13].

The technology of semiautomatic extraction of structural elements from mathematical texts is proposed in [12, 14]. Using this technology, the source text of the mathematical document is annotated using the sTeX macro command. Another technology for presenting a mathematical document is developed on the basis of OMDoc [15].

One should also note that some efforts have been made to use keyword search to extract the mathematical content of documents. For example, the WolframAlpha computational engine (https://www.wolframalpha.com/) can handle queries by keywords. However, this mechanism does not provide similar functionality for documents from scientific collections.

MCAT Math Retrieval System [16, 17] uses the SVM classifier to detect descriptions of mathematical expressions and extends the base line of the TF-IDF ranking to find formulas in documents presented in MathML format. Note that unlike this tool, our solution which is mentioned above, is more flexible as it allows lexical character values in terms of the OntoMathPRO ontology and, therefore, allows the use of ISA type relationships. In addition, our search interface supports filtering by the context of the document structure, that is by a certain segment of the document that contains the corresponding formula (for example, by theorem or definition).

An overview of search methods in mathematical documents, including mathematical formulas, is given in [2].

To conclude this section, we must mention the search methods we developed earlier. They are based on the use of MathML-representation of mathematical documents [11]. Our algorithm converts formulas from TeX-format to MathML format. We created a prototype of an information retrieval system for a collection that includes articles from the Lobachevskii Journal of Mathematics. For the end user, the query input interface supports convenient syntax. Search results include selected occurrences of formulas, as well as document metadata.

III. OntoMath Ecosystem

OntoMath is a digital ecosystem of ontologies developed by us, as well as of text analytics tools and mathematical knowledge management applications [18]. This system consists of the following components:

- ontology of structural elements of mathematical scientific works called Mocassin [19, 20];

- ontology of the concepts of mathematical knowledge OntoMathPRO [18, 21, 22];

- semantic publishing platform [23];

- semantic search service by formulas [18];

- recommender system for forming lists of related publications [24].

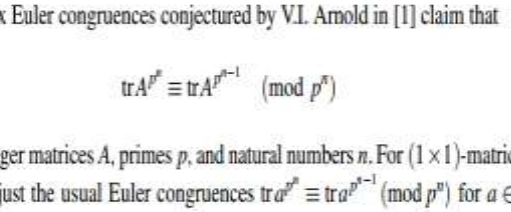
The main component of the OntoMath ecosystem is the semantic publishing platform. It creates an LOD representation for a set of mathematical articles in LaTeX. The generated mathematical data set includes metadata, logical structure of documents, terminology and mathematical formulas. Article metadata, logical structure of documents and terminology are expressed in terms of ontologies AKT Portal, Mocassin and OntoMathPRO, respectively. The ontology Mocassin, in turn, is based on Semantically Annotated LaTeX (SALT) Document Ontology, which is the ontology of the rhetorical structure of scientific publications [25]. Ontologies Mocassin and OntoMathPRO are parts of the OntoMath ecosystem, but SALT is an outside ontology. Two applications are built using the semantic publication platform: the semantic search service for formulas and the recommender system.

IV. OntoMath Semantic Formula Search

The novelty of the semantic approach we realized consists in expanding the context of the mathematical formula in search organization. The extended context of the formula includes the type of the structural element (section, theorem, etc.), as well as text fragments defining relations "terms – conditional symbols" and "symbols – formulas". The first relation is a textual definition of the meaning of a symbol in the context of a term. The second relation indicates the relationship between the symbol and the formula.

Using the information in the extended context, we can find all the formulas containing the given variable. For example, we can implement a query: find a formula that contains the variable "matrix norm". The corresponding search allows, on the one hand, to more accurately take into account the semantics of the variables in the formula and, on the other hand, provide the user with the corresponding fragment of the document. As a rule, localization of the formula (that is, the possibility of more precise determination of its position in documents) is also not provided by the search systems described above.

Let us consider the example of the extended context of the formula. Fig. 1 shows a fragment of a mathematical paper.


The matrix Euler congruences conjectured by V.I. Arnold in [1] claim that

$$\text{tr}A^{p^n} \equiv \text{tr}A^{p^{n-1}} \pmod{p^n} \quad (1)$$

for all integer matrices A, primes p, and natural numbers n. For (1×1)-matrices, these are just the usual Euler congruences $\text{tr}\,a^{p^n} \equiv \text{tr}\,a^{p^{n-1}} \pmod{p^n}$ for $a \in \mathbb{Z}$.

Figure 1. Fragment of a mathematical paper

Variables *A*, *p* and *n* for formula (1) are defined in the context after the formula. The result of the binding of the variables included in this formula is shown in Fig. 2. The binding algorithm reveals the meaning of the variables and implements the parsing of the context. The binding algorithm uses an empirical approach to estimate the distance between a symbol and its text content, and also processes annotations

created by NLP (Natural Language Processing) for mathematical work. Some variables included in the formula are related to the concepts of the OntoMathPRO ontology. Thus, we establish the connection between the concepts from the ontology and the variables in the formula.

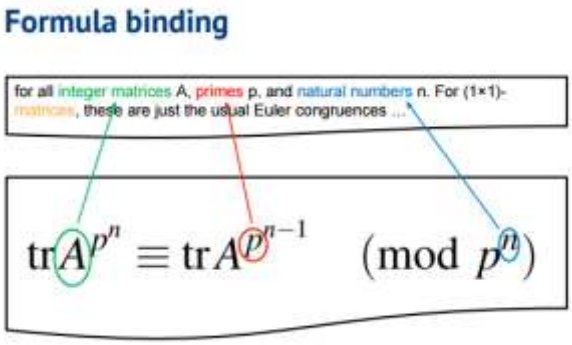


Figure 2. Binding formula variables

A variable in a formula is a symbol that designates a mathematical object. Mathematical symbols can denote numbers (constants), variables, operations, functions, punctuation, grouping, and other aspects of logical syntax. Specific branches and applications of mathematics usually have special naming conventions for variables. However, in some formulas, non-standard variable names can be used. The OntoMath Formula Search Engine tool allows one to find mathematical formulas containing a given mathematical object, regardless of the name of the variable included in this object. For example, if we want to find a formula containing a mathematical object (for example, curvature), the search engine will find all the formulas that include this object (even with different names for the variable).

Using the output, this system can find formulas that contain not only this object, but also objects located lower in the ontology hierarchy. For example, to search for formulas containing a polygon, the OntoMath Formula Search system can find formulas that contain not only polygons, but also other objects in the hierarchy (for example, triangles, parallelograms, trapezoids, hexagons, and others). OntoMath also allows you to restrict the search by the formula in the part (area) of the document that you specify. For example, you can search only in certain sections of the document or in a specific part of the theorem.

The search functions of the OntoMath Formula Engine system described above are different from popular search services such as Sprinter LaTeX search, Wikipedia formulas search, or Wolfram formula search system. These services have great potential and good properties, including their stability for renaming variables and converting expressions. However, they are syntactical and allow you to find only formulas that contain a given template.

We have implemented two applications for finding mathematical formulas, such as syntactic search for formulas in MathML and search for semantic ontologies. The syntactic search uses formula descriptions from documents converted to tex-format. Semantic publishing platform builds semantic representation for a collection of mathematical papers in LaTeX, as RDF dataset (Figure 3). Semantic Search UI is a single-page web application, that works on top of generated dataset via SPARQL endpoint. Figure 4 shows the results of searching for formulas containing a given mathematical object (*ring*) from OntoMath^{PRO} ontology. In this case, the formula variables can use any symbolic notations.

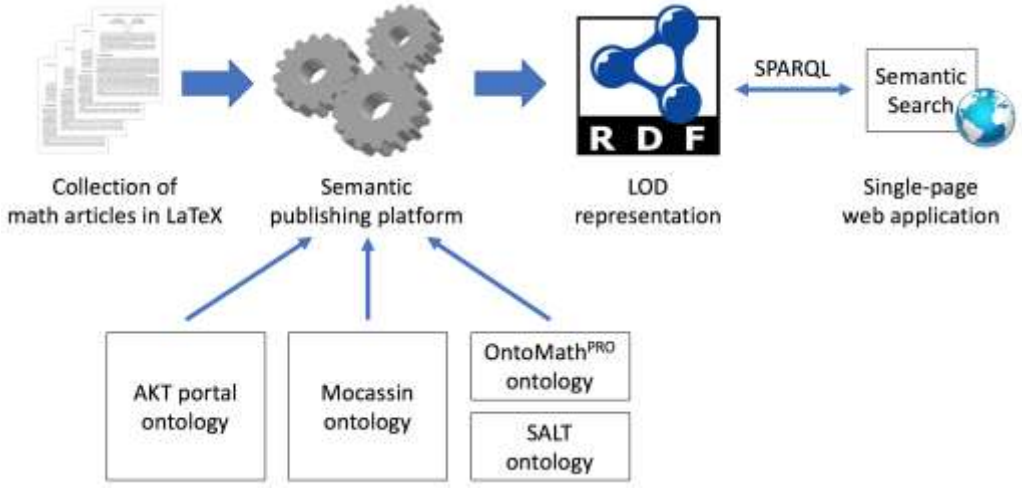


Figure 3. OntoMath^{PRO} Formula Search Workflow

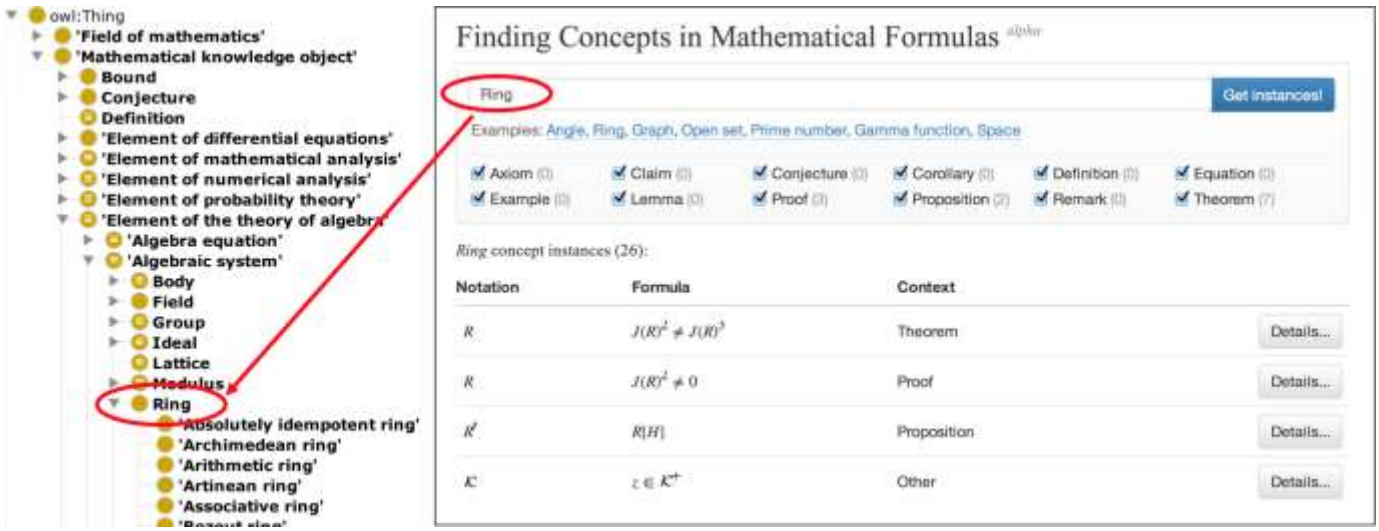


Figure 4. OntoMath^{PRO} Search UI

V. LOBACHEVSKII DML

Lobachevskii Digital Mathematics Library (Lobachevskii-DML, http: // www. Lobachevskii-dml.ru/) is developed by us in accordance with the basic principles of the WDML project. The main task is to develop tools for managing mathematical content, taking into account not only the specifics of mathematical documents, but also the features of processing Russian-language texts. A particular task of creating this electronic library is integrating mathematical resources of Kazan University. The described semantic search service is available on Lobachevskii DML website: http://lobachevskii-dml.ru:8890/mathsearch/.

ACKNOWLEDGMENT

The research we are carrying out is directed towards further development of methods and tools of mathematical knowledge management. We have been working in this direction since 1998, with the support of Russian Foundation for Basic Research (RFBR), Kazan Federal University (KFU) and the Academy of Sciences of the Republic of Tatarstan.

This work was financed by Kazan Federal University for the state task in the field of scientific activity (grant agreement No. 1.2368.2017), as well as RFBR and Government of the Republic of Tatarstan in the framework of scientific projects No. 15-07-08522, 15-47- 02472.

REFERENCES

[1] M. Líška. "Building the Ultimate Math Search Engine," Ph.D. Thesis Proposal, Masaryk University, Faculty of Informatics, Brno, 97 p., 2015.

[2] F. Guidi, C. Sacerdoti Coen, "A Survey on Retrieval of Mathematical Knowledge," Math.Comput.Sci., no.10, pp. 409–427, 2016, doi:10.1007/s11786-016-0274-0.

[3] M.Q Shatnawi., Q.Q. Abuein. "A Digital Ecosystem-based Framework for Math Search Systems," International Journal of Advanced Computer Science and Applications, vol. 3, no. 3, pp. 78–83, 2012, doi:10.14569/IJACSA.2012.030314.

[4] P. Sojka, M. Líška. "Indexing and Searching Mathematics in Digital Libraries," In: Davenport J.H., Farmer W.M., Urban J., Rabe F. (eds) Intelligent Computer Mathematics. CICM 2011. Lecture Notes in Computer Science, vol 6824. Springer, Berlin, Heidelberg, pp. 228-243, 2011, doi:10.1007/978-3-642-22673-1_16.

[5] T. Bouche. "Reviving the Free Public Scientific Library in the Digital Age? The EuDML project," In: Kaiser K., Krantz S., Wegner B. (Eds.) Topics and Issues in Electronic Publishing, JMM, Special Session, San Diego, pp. 57-80, 2013.

[6] A. Elizarov, A. Kirillovich, E. Lipachev, and O. Nevzorova. "Digital Ecosystem OntoMath: Mathematical Knowledge Analytics and Management," Communications in Computer and Information Science, Springer, vol. 706, pp 33-46, 2017, doi:10.1007/978-3-319-57135-5_3.

[7] T.W. Cole, I. Daubechies, K.M. Carley, J.L. Klavans, Y. LeCun, M. Lesk, C.A. Lynch, P. Olver, J. Pitman, and Z.J. Xia. "Developing a 21st century global library for mathematics research. Washington, D.C.," The National Academies Press, Washington, D.C, 2014.

[8] P.J. Olver, "The World Digital Mathematics Library: report of a panel discussion," Proceedings of the International Congress of Mathematicians, August 13–21, 2014, Seoul, Korea. Kyung Moon SA, vol. 1, pp. 773–785, 2014.

[9] R. Miller, and A. Youssef. "Augmenting Presentation MathML for Search.," In: Autexier S. et al. (Eds.): AISC/Calculemus/MKM 2008, LNAI 5144, pp. 536–542, 2008.

[10] R. Zanibbi, A. Orakwue. "Math Search for the Masses: Multimodal Search Interfaces and Appearance-Based Retrieval," In: Kerber M., Carette J., Kaliszyk C., Rabe F., Sorge V. (eds) Intelligent Computer Mathematics. CICM 2015. Lecture Notes in Computer Science, vol 9150. Springer, Cham, pp. 18–36, 2015, doi:10.1007/978-3-319-20615-8_2.

[11] A.M. Elizarov, E.K. Lipachev, M.A. Malakhaltsev, "Web Technologies for Mathematicians: The Basics of MathML. A Practical Guide," Fizmatlit, Moscow, 2010 (In Russian).

[12] M. Kohlhase, B.A. Matican, CC. Prodescu. "MathWebSearch 0.5: Scaling an Open Formula Search Engine," In: Jeuring J. et al. (eds) Intelligent Computer Mathematics. CICM 2012. Lecture Notes in Computer Science, vol 7362. Springer, Berlin, Heidelberg, pp. 342-357, 2012, doi:10.1007/978-3-642-31374-5_23.

[13] P. Sojka, M. Líška. "The art of mathematics retrieval," Proceedings of the 11th ACM symposium on Document engineering. – ACM, pp. 57-60, 2011.

[14] A. Kohlhase, M. Kohlhase and C. Lange "sTeX – a system for flexible formalization of linked data," Proceedings of the 6th International Conference on Semantic Systems, ACM, pp. 57-60, 2010, doi: 10.1145/2034691.2034703.

[15] M. Kohlhase. “OMDoc-An Open Markup Format for Mathematical Documents [version 1.2],” Lecture Notes in Computer Science, vol. 4180, Springer, 428 p, 2006, doi:10.1007/11826095.

[16] G. Topić, G.Y. Kristianto, M.-Q. Nghiem, A. Aizawa. “The MCAT Math Retrieval System for NTCIR-10 Math Track,” In: Proceedings of the 10th NTCIR Conference, pp. 680-685, 2013.

[17] G.Y. Kristianto, G. Topić, F. Ho, A. Aizawa. “The MCAT Math Retrieval System for NTCIR-11 Math Track,” In: Proceedings of the 11th NTCIR Conference, pp. 120-126, 2014.

[18] O. Nevzorova, N. Zhiltsov, A. Kirillovich, and E. Lipachev, “OntoMathPRO ontology: a linked data hub for mathematics,” In: Klinov P., Mouromtsev D. (eds.) KESW 2014. Communications in Computer and Information Science, Springer, vol. 468, pp. 105–119, 2014, doi: 10.1007/978-3-319-11716-4_9.

[19] V. Solovyev, N. Zhiltsov, “Logical structure analysis of scientific publications in mathematics,” In: Akerkar, R. (ed.) Proceedings of the International Conference on Web Intelligence, Mining and Semantics (WIMS 2011), ACM DL, vol. 21, pp. 1–9, 2011, doi: 10.1145/1988688.1988713.

[20] O.A. Nevzorova, E.V. Birialtsev, and N.G. Zhiltsov, “Mathematical Text Collections: Annotation and Application for Search Tasks,” Sci. Tech.Inf. Proc., vol. 40, no. 6, pp. 386–395, 2013

[21] A. Elizarov, A. Kirillovich, E. Lipachev, O. Nevzorova, V. Solovyev, and N. Zhiltsov, “Mathematical knowledge representation: semantic models and formalisms,” Lobachevskii J. of Mathematics, vol. 35, no 4, pp. 347–353, 2014, doi:10.1134/S1995080214040143.

[22] A.M. Elizarov, E.K. Lipachev, O.A. Nevzorova, and V.D. Solov’ev, “Methods and means for semantic structuring of electronic mathematical documents,” Doklady Mathematics, vol. 90, no. 1, pp. 521–524, 2014, doi: 10.1134/S1064562414050275.

[23] O. Nevzorova, N. Zhiltsov, D. Zaikin, O. Zhibrik, A. Kirillovich, V. Nevzorov, and E. Birialtsev, “Bringing Math to LOD: a semantic publishing platform prototype for scientific collections in mathematics,” In: Alani H. et al (eds) 12th Int. Semantic Web Con-ference, Sydney, NSW, Australia, October 21–25, 2013, Proceedings, Part I. Lecture Notes in Computer Science, vol. 8218, pp. 379–394. Springer Berlin Heidelberg, 2013.

[24] A.M. Elizarov, A.B. Kirillovich, E.K. Lipachev, A.B. Zhizhchenko, and N.G. Zhil'tsov, “Mathematical Knowledge Ontologies and Recommender Systems for Collections of Documents in Physics and Mathematics,” Doklady Mathematics, vol. 93, no. 2, pp. 231–233, 2016, doi:10.1134/S1064562416020174.

[25] T. Groza, S. Handschuh, K. Möller, S. Decker. “SALT - Semantically Annotated LaTeX for Scientific Publications,”. In: Franconi E., Kifer M., May W. (eds) The Semantic Web: Research and Applications. ESWC 2007. Lecture Notes in Computer Science, vol 4519. Springer, Berlin, Heidelberg, pp. 518-532, 2007, doi: 10.1007/978-3-540-72667-8_37.