WebMIaS on Docker*

Math-Aware Search in a Single Line of Code

Dávid Lupták[0000−0001−9600−7597], Vít Novotný[0000−0002−3303−4130], Michal Štefánik[0000−0003−1766−5538], and Petr Sojka[0000−0002−5768−4007]

Faculty of Informatics, Masaryk University
{dluptak,witiko,stefanik.m}@mail.muni.cz, sojka@fi.muni.cz
https://mir.fi.muni.cz/

Abstract. Math informational retrieval (MIR) search engines are absent in the wide-spread production use, even though documents in the STEM fields contain many mathematical formulae, which are often more important than text for understanding. We have developed and open-sourced the WebMIaS MIR system that has been successfully deployed in the European Digital Mathematical Library (EuDML). However, its deployment is difficult to automate, and the solutions developed so far to tackle this challenge are imperfect in terms of speed, maintenance, and robustness. In this paper, we will describe the virtualization of WebMIaS using Docker that solves all three problems and allows anyone to deploy WebMIaS in a single line of code. The publicly available Docker image will also help the community push the development of math-aware search engines in the ARQMath workshop series.

Keywords: Math Information Retrieval · WebMIaS · MIaS · Docker Virtualization · Digital Mathematical Libraries · Math Web Search · EuDML · ARQMath.

1 Introduction

Searching for math formulae does not appear as a task for search engines at first glance. Text retrieval is dominant among search engines, while math-awareness is a specialized area in the field. Springer's LaTeX Search, MathWebSearch system of zbMATH Open (formerly known as Zentralblatt MATH), MIaS (Math Indexer and Searcher) of European Digital Mathematical Library (EuDML) are examples of a few systems deployed in the production use with math search functionality. We have developed and open-sourced MIaS system [9] running on the industry-leading, robust, and highly-scalable full-text search engine Apache Lucene with our own preprocessing workflow for math formulae.

MIaS processes text and math separately. The text is tokenized and stemmed to unify inflected word forms. Math is expected to be in the MathML format, then canonicalized, ordered, tokenized, and unified (see Figure 1).

To provide a web user interface to MIaS, we have developed and open-sourced WebMIaS [9,6]. Users can input their query in a combination of text, and math

* The second author's work was graciously funded by the South Moravian Centre for International Mobility as a part of the Brno Ph.D. Talent project.

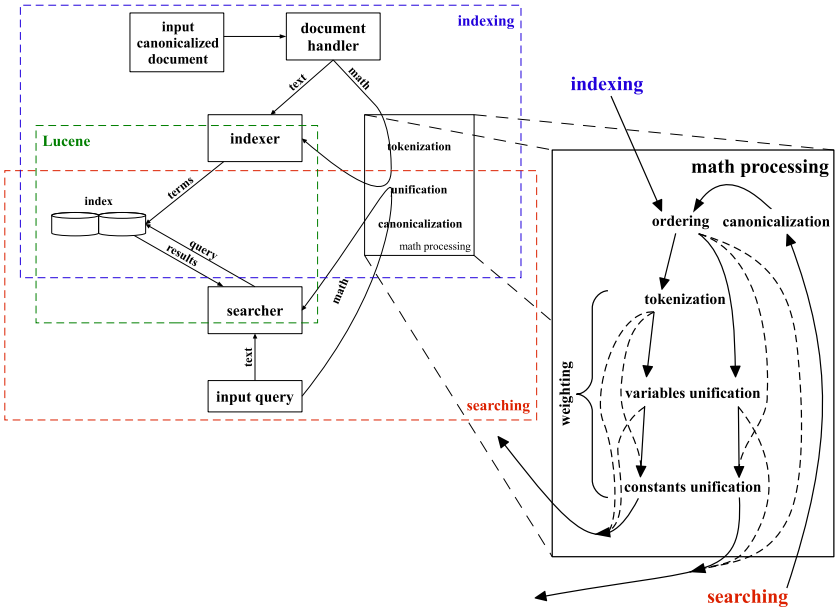

Fig. 1. MIaS ecosystem architecture with indexing and searching phases overlapping over Lucene index; math processing elaborated as an integral part of both phases.

with a native support for LaTeX or MathML. Matches are conveniently highlighted in the search results. The user interface of WebMIaS is shown in Figure 2.

Although the (Web)MIaS system has been deployed in the digital mathematics library (DML) already, a complicated deployment process might be an issue for offering it for other DMLs. To solve this problem, we will describe the virtualization of WebMIaS using Docker [3] that allows anyone to deploy WebMIaS in a single line of code. Whether you have a DSpace system, another open access repository, or just a bunch of MathML documents at your hands, you can try WebMIaS out. Otherwise, we provide the MREC dataset [4] to play around with.

In the rest of this paper, we describe a deployment process in Section 2, present multiple evaluations in Section 3, and conclude in Section 4.

2 Deployment process description

All system modules in the MIaS ecosystem are Java projects, so 1) Java environment prerequisites need to be installed beforehand, then we can 2) build the respective system modules. The next step in the workflow is 3) to index a dataset – MathML documents – using the command-line interface of MIaS. Finally, we can 4) run Apache Tomcat with WebMIaS servlet as a user interface.

Over the years, we have attempted to automate the above steps into running a single `make` of Makefile or Jupyter Notebook. However, these imperfect solutions were slow, fragile, and hard to maintain. We propose a better solution using lightweight virtualization via Docker with instant deployment, short yet

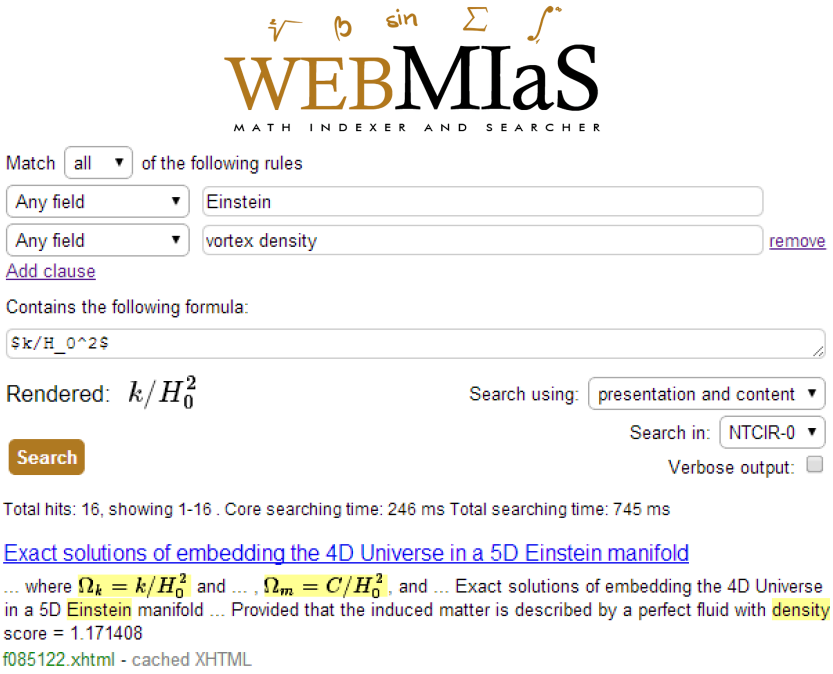

Fig. 2. Searching text and formulae in a single query with WebMIaS user interface.

powerful Dockerfile configuration, and a complete workflow that automates all steps. Moreover, GitHub Actions provide continuous integration and automatic publishing of Docker images to Docker Hub.

Both MIaS and WebMIaS are containerized into separate docker images, named `miratmu/mias` and `miratmu/webmias`, respectively. This allows users to run both indexing and retrieval without a specific configuration of the environment. Resolving dependencies and building all system modules is up to the continuous integration workflow (see Figure 3), and users get only Docker containers with everything prebuilt. After downloading a dataset to the working directory, users can index `dataset` directory using MIaS (`index` directory for storing the index),

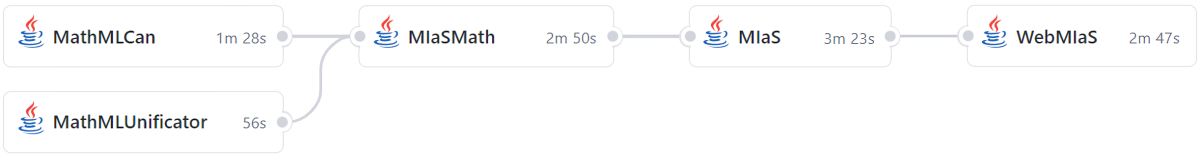

Fig. 3. Continuous integration of *WebMIaS*: *MathMLCan* canonicalizes different MathML encodings of equivalent formulae. *MathMLUnificator* generalizes distinct mathematical formulae so that they can be structurally unified. *MIaSMath* adds math processing capabilities to Lucene or Solr. *MIaS* indexes text with math in Lucene/Solr-based full-text search engines. Finally, *WebMIaS* provides a web interface for *MIaS*.

```
1 $ wget https://mir.fi.muni.cz/MREC/MREC2011.4.439.tar.bz2
2 $ mkdir dataset ; tar xj -f MREC2011.4.439.tar.bz2 -C dataset
3 $ docker run -v "$PWD"/dataset:/dataset:ro -v "$PWD"/index:/index:rw --rm
  ↪ miratmu/mias
4 $ docker run -v "$PWD"/dataset:/dataset:ro -v "$PWD"/index:/index:ro --rm
  ↪ --name webmias -d -p 127.0.0.1:8888:8080 miratmu/webmias
```

Listing 1: Downloading and indexing the MREC2011.4 dataset for WebMIaS (lines 1–3), and deploying WebMIaS in a single line (n. 4) of code.

see Listing 1. Finally, users can deploy WebMIaS in a single line of code with specified `dataset` and `index` directories, assigned name `webmias`, and detached port `8888` on the `localhost`. The WebMIaS system is then running at http://localhost:8888/WebMIaS. For detailed technical insight into the Docker run command, please see its documentation[1].

3 Evaluation

We performed a speed evaluation of MIaS on the MREC dataset [4] (see Table 1), and a quality evaluation on the NTCIR-10 Math [1,5], NTCIR-11 Math-2 [2,9] (see Table 2), NTCIR-12 MathIR [10,8], and ARQMath 2020 [11,7] datasets. We also measured the time to deploy WebMIaS without Docker (see Figure 3).

Speed evaluation shows that the indexing time of our system is linear in the number of indexed documents and that the average query time is 469 ms. Additionally, the dockerization of WebMIaS reduces the deployment time from 10 minutes and 28 seconds to a matter of seconds. With respect to quality evaluation, MIaS has notably won the NTCIR-11 Math-2 task.

4 Conclusion

An open-source environment brings reproducibility and the possibility of trying out the projects of one's interest, usually without limitations. Installation and following build instructions are often hard to follow with many prerequisites and possible conflicts with the running operating environment on the go. Automation tools, continuous integration, and package virtualization ease the development process. With stated motivation and hopes to please the math community, we

[1] https://docs.docker.com/engine/reference/run/

Table 1. Linearity of indexing speed on the MREC dataset using 448G of RAM, and eight Intel Xeon™ X7560 2.26 GHz CPUs.

	Mathematical (sub)formulae		Indexing time (min)	
Docs	Input	Indexed	Real	CPU
10,000	3,406,068	64,008,762	35.75	35.05
100,000	36,328,126	670,335,243	384.44	366.54
439,423	158,106,118	2,910,314,146	1,747.16	1,623.22

Table 2. Quality evaluation results on the NTCIR-11 Math-2 dataset. The mean average precision (MAP), and precisions at ten (P@10), and five (P@5) are reported for queries formulated using Presentation (PMath), and Content MathML (CMath), a combination of both (PCMath), and LaTeX. Two different relevance judgement levels of ≥ 1 (partially relevant), and ≥ 3 (relevant) were used to compute the measures. Number between slashes (/·/) is our rank among all teams.

Measure	Level	PMath	CMath	PCMath	LaTeX
MAP	3	0.3073	**0.3630 /1/**	0.3594	0.3357
P@10	3	0.3040	**0.3520 /1/**	0.3480	0.3380
P@5	3	0.5120	**0.5680 /1/**	0.5560	0.5400
P@10	1	0.5020	0.5440	**0.5520 /1/**	0.5400

have dockerized our math-aware web search interface WebMIaS. One can do that easily in a single line of code now. The software is accessible and at the fingertips of the math community (see https://github.com/MIR-MU/WebMIaS).

References

1. Aizawa, A., Kohlhase, M., Ounis, I.: NTCIR-10 Math Pilot Task Overview. In: Proc. of the 10th NTCIR Conference. pp. 654–661. NII, Tokyo, Japan (2013)
2. Aizawa, A., Kohlhase, M., Ounis, I., Schubotz, M.: NTCIR-11 Math-2 Task Overview. In: Proc. of the 11th NTCIR Conference. pp. 88–98. NII, Tokyo (2014), http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings11/pdf/NTCIR/OVERVIEW/01-NTCIR11-OV-MATH-AizawaA.pdf
3. Boettiger, C.: An introduction to Docker for reproducible research. ACM SIGOPS Operating Systems Review **49**(1), 71–79 (2015)
4. Líška, M., Sojka, P., Růžička, M., Mravec, P.: Web Interface and Collection for Mathematical Retrieval: WebMIaS and MREC. In: Proc. of DML 2011 workshop. pp. 77–84 (2011), https://hdl.handle.net/10338.dmlcz/702604
5. Líška, M., Sojka, P., Růžička, M.: Similarity Search for Mathematics: Masaryk University team at the NTCIR-10 Math Task. In: Proc. of the 10th NTCIR Conference. pp. 686–691. NII, Tokyo, Tokyo (2013), https://research.nii.ac.jp/ntcir/workshop/OnlineProceedings10/pdf/NTCIR/MATH/06-NTCIR10-MATH-LiskaM.pdf
6. Líška, M., Sojka, P., Růžička, M.: Math indexer and searcher web interface: Towards fulfillment of mathematicians' information needs. In: Proc. of CICM 2014. pp. 444–448. Springer, Zurich (2014), https://doi.org/10.1007/978-3-319-08434-3_36
7. Novotný, V., Sojka, P., Štefánik, M., Lupták, D.: Three is better than one. In: CEUR Workshop Proceedings. pp. 1–30. Thessaloniki, Greece (2020), http://ceur-ws.org/Vol-2696/paper_235.pdf
8. Růžička, M., Sojka, P., Líška, M.: Math Indexer and Searcher under the Hood: Fine-tuning Query Expansion and Unification Strategies. In: Proc. of the 12th NTCIR Conference. pp. 331–337. NII Tokyo (2016), https://research.nii.ac.jp/ntcir/workshop/OnlineProceedings12/pdf/ntcir/MathIR/05-NTCIR12-MathIR-RuzickaM.pdf
9. Růžička, M., Sojka, P., Líška, M.: Math Indexer and Searcher under the Hood: History and Development of a Winning Strategy. In: Proc. of the 11th NTCIR Conference. pp. 127–134 (2014), https://is.muni.cz/auth/publication/1201956/en
10. Zanibbi, R., Aizawa, A., Kohlhase, M., Ounis, I., Topic, G., Davila, K.: NTCIR-12 MathIR task overview. In: Proc. of the 12th NTCIR. pp. 299–308. NII Tokyo (2016)
11. Zanibbi, R., Oard, D.W., Agarwal, A., Mansouri, B.: Overview of ARQMath 2020: CLEF lab on answer retrieval for questions on math. In: Proc. of Int. Conf. CLEF 2020. pp. 169–193. Springer (2020)