Методы Big Math и интеграция математических знаний

Елизаров Александр Михайлович[1] amelizarov@gmail.com
Липачев Евгений Константинович[1]⋆ elipachev@gmail.com
[1]Казань, Казанский (Приволжский) федеральный университет

Термин “Big Data”, широко используемый в настоящее время в различных предметных областях, применительно к математике требует определенных уточнений: в математике все данные существенны, кроме того, в математических документах многие их части, особенно формулы, являются своеобразным кодом, требующим расшифровки и специального толкования. Далее, при решении математических задач являются существенно большими ожидания от использования ИКТ. Здесь можно провести аналогию с тем, как вычислительные машины полностью устранили ручные вычисления. Вычисления всегда требовали применения особых методов и нестандартных организационных решений, позволяющих справиться с объемом (Volume – одна из характеристик больших данных) и преодолеть барьер вычислительных возможностей отдельного человека. Если говорить о Velocity как одной из характеристик больших данных, то длительность ручных вычислений иллюстрирует пример вычисления числа Пи: В. Шенкс (William Shanks, 1873 г.) потратил 15 лет на вычисление 707 знаков этого числа (однако только 555 из них оказались верными). Помимо вычислений и подготовки документов необходимы инструменты интеллектуального поиска, в том числе, рекомендательные системы для нахождения научных статей, близких по содержанию; сервисы терминологического аннотирования; персональные информационные помощники и цифровые платформы для автоматизации издательской деятельности.

Недавно J. Carette, W.M. Farmer, M. Kohlhase и F. Rabe (arXiv:1904.10405v1 [cs.MS] 23 April 2019) предложили использовать, по аналогии с Big Data, термин Big Math для обозначения области создания методов и разработки программных систем поддержки математических исследований. Ими выделены пять основных аспектов Big Math:

• Inference (вывод утверждений путем дедукции),

• Computation (алгоритмическое преобразование представлений математических объектов в формы, более легкие для понимания),

• Tabulation (создание статических, конкретных данных, относящихся к математическим объектам и структурам, которые можно легко хранить, запрашивать и совместно использовать),

• Narration (приведение результатов в форму, которая может быть усвоена людьми),

• Organization (модульная организация математических знаний).

Основная задача математических программных систем заключается сегодня в интеграции указанных аспектов, составляющих Big Math. Система цифровых математических библиотек, создаваемая в настоящее время, призвана консоли-

дировать и сделать доступными как современные математические знания, так и математические знания, содержащиеся в документах, созданных в доцифровой период. Для достижения этой цели в рамках цифровых библиотек разрабатываются методы управления цифровой информацией, учитывающие особенности представления математического контента.

В области интеграции математических знаний наиболее значительными являются инициатива Global Digital Mathematics Library (GDML, https://doi.org/10.1007/978-3-319-62075-6_5) и проект World Digital Mathematics Library (WDML, https://arxiv.org/ftp/arxiv/papers/1404/1404.1905.pdf). Его основная задача – объединение в распределенной системе электронных коллекций всего корпуса цифровых математических документов. На интеграцию европейских математических ресурсов направлен проект The European Digital Mathematics Library (EuDML, https://initiative.eudml.org/). Этот проект рассматривается как один из этапов построения WDML.

В соответствии с основными принципами WDML в Казанском университете создается цифровая математическая библиотека Lobachevskii Digital Mathematics Library (Lobachevskii-DML, https://lobachevskii-dml.ru/). Построение этой библиотеки предполагает разработку инструментов управления математическим контентом, учитывающих не только специфику математических текстов, но и особенности обработки русскоязычных текстов. Еще одной задачей этой цифровой библиотеки является интеграция математических ресурсов Казанского университета и их включение в глобальную научную инфраструктуру, в частности, MathNet.Ru и EuDML.

В исследованиях, выполненных нашей группой (см. [1]), разработаны подходы к управлению большими коллекциями цифровых математических документов, основанные на семантических методах и согласующиеся с принципами WDML, а также относящиеся к направлениям, составляющим Big Math. Эти подходы развиваются и уже частично практически реализованы в цифровой математической библиотеке Lobachevskii-DML. Предложены методы формирования цифровых коллекций из набора документов – научных статей, монографий, докладов, представленных в различных форматах хранения. На основе анализа структуры документов и стилевых особенностей их оформления разработан алгоритм экстракции их метаданных. Создан программный инструмент разделения сборников статей на отдельные документы и формирования их семантического представления. На примере «Трудов Математического центра им. Н.И. Лобачевского», имеющих различные формат и структуру, реализован алгоритм создания цифровой коллекции и ее включения в Lobachevskii-DML.

Разработаны: алгоритмы пополнения электронных коллекций цифровой библиотеки Lobachevskii-DML и формирования метаданных документов этих коллекций в выбранных форматах; сервисы нормализации этих метаданных в соответствии с DTD-правилами и XML-схемами NISO JATS и DBLP; алгоритмы

создания обязательного и фундаментального наборов метаданных коллекций в соответствии с правилами EuDML.

Работа выполнена в рамках программы развития Регионального научно-образовательного математического центра Приволжского федерального округа, соглашение № 075-02-2020-1478/1.

[1] *Elizarov A. M. and Lipachev E. K.* Big Math Methods in Lobachevskii–DML Digital Library // Data Analytics and Management in Data Intensive Domains, 2019. Pp. 59–72.

Big Math Methods and Mathematical Knowledge Integration

Alexander Elizarov[1] amelizarov@gmail.com
Evgeny Lipachev[1]⋆ elipachev@gmail.com
[1] Kazan, Kazan (Volga Region) Federal University

The term "Big Data", which is currently widely used in various subject areas, in relation to mathematics requires certain clarifications: in mathematics, all data is essential, in addition, in mathematical documents, many of their parts, especially formulas, are a kind of code that requires decoding and special interpretation. Further, when solving mathematical problems, expectations from the use of ICT are significantly higher. An analogy can be drawn here with how computers completely eliminated manual computation. Computing has always required the use of special methods and non-standard organizational solutions to cope with volume (Volume is one of the characteristics of big data) and overcome the barrier of the computational capabilities of an individual. If we talk about Velocity as one of the characteristics of big data, then the duration of manual calculations illustrates an example of calculating the number Pi: W. Shanks (William Shanks, 1873) spent 15 years calculating 707 digits of this number (however, only 555 of them turned out to be correct). In addition to calculations and preparation of documents, intelligent search tools are needed, including recommendation systems for finding scientific articles that are similar in content; terminological annotation services; personal information assistants and digital platforms for publishing automation.

J. Carette, W.M. Farmer, M. Kohlhase and F. Rabe (arXiv: 1904.10405v1 [cs.MS] 23 April 2019) proposed to use, by analogy with the term Big Data, the term Big Math to denote the field of creating methods and developing software systems to support mathematical research. They highlighted 5 main aspects of Big Math:

- Inference (output of statements by deduction);
- Computation (algorithmic transformation of representations of mathematical objects into forms that are easier to understand);
- Tabulation (creating static, specific data related to mathematical objects and structures that can be easily stored, queried and shared);
- Narration (bringing the results into a form that people can assimilate);
- Organization (modular organization of mathematical knowledge).

The main task of mathematical software systems today is to integrate the aspects that make up Big Math. The system of digital mathematical libraries currently being created is intended to consolidate and make accessible both modern mathematical knowledge and the knowledge contained in articles and books published in the pre-digital period. To achieve this goal, in the framework of digital libraries, methods for managing digital information are developed that take into account the characteristics of the presentation of mathematical content.

In the area of integrating mathematical knowledge, the most significant are the Global Digital Mathematics Library initiative (GDML, https://doi.org/10.1007/978-3-319-62075-6_5) and the World Digital Mathematics Library project (WDML, https://arxiv. org/ftp/arxiv/papers/1404/1404.1905.pdf). Its main task is to unite the entire corpus of digital mathematical documents in a distributed system of electronic collections. The European Digital Mathematics Library project (EuDML, https://initiative.eudml.org/) aims to integrate European mathematical resources. This project is considered as one of the stages of building WDML.

In accordance with the basic principles of WDML, a digital library Lobachevskii Digital Mathematics Library (Lobachevskii-DML, https://lobachevskii-dml.ru/) is being created at the Kazan University. The construction of this library involves the development of management tools for mathematical content that take into account not only the specifics of mathematical texts, but also the peculiarities of processing Russian-language texts. Another objective of this digital library is the integration of the mathematical resources of Kazan University and their inclusion in the global scientific infrastructure, in particular, MathNet.Ru and EuDML.

In the studies carried out by our group (see [1]), approaches to managing large collections of digital mathematical documents based on semantic methods and consistent with the WDML principles, as well as related to the directions that make up Big Math, have been developed. These approaches are being developed and already partially practically implemented in the digital mathematical library Lobachevskii-DML. Methods for the formation of digital collections from a set of documents – scientific articles, monographs, reports, presented in various storage formats are proposed. Based on the analysis of the structure of documents and the style features of their design, an algorithm for extracting their metadata has been developed. A software tool has been created for dividing collections of articles into separate documents and forming their semantic representation. On the example of "Proceedings of the N.I. Lobachevskii Mathematical Center", which have different formats and structures, an algorithm for creating a digital collection and its inclusion in Lobachevskii-DML was implemented.

We have developed: algorithms for replenishing electronic collections of the digital library Lobachevskii-DML and forming metadata for documents of these collections in selected formats; services for normalizing this metadata in accordance with DTD rules and NISO JATS and DBLP XML schemas; algorithms for creating mandatory and fundamental sets of collection metadata in accordance with EuDML rules.

The work was carried out within the framework of the development program of the Regional Scientific and Educational Mathematical Center of the Volga Federal District, agreement No. 075-02-2020-1478/1.

[1] *Elizarov A. M. and Lipachev E. K.* Big Math Methods in Lobachevskii–DML Digital Library // Data Analytics and Management in Data Intensive Domains, 2019. Pp. 59–72.