Moritz SCHUBOTZ, Konstanz

Mathematische Formeln in Wikipedia

Mathematical Information Retrieval

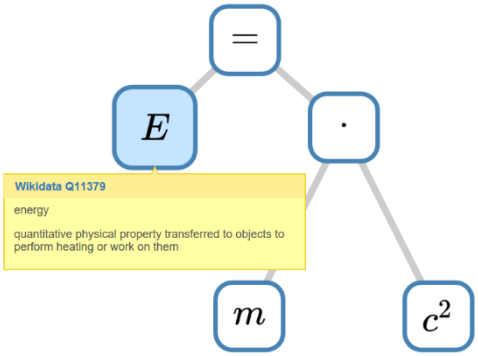

Abb. 1: Semantische Anreicherung des Bezeichners E in der Energie-Masse-Äquivalenz (Schubotz, 2017c)

Mathematical Information Retrieval (MathIR) ist ein Forschungsbereich an der Schnittstelle zwischen Mathematik, Informatik und Linguistik, der sich mit der computergestützten Verwaltung und Verarbeitung von mathematischem Wissen beschäftigt. In vielen wissenschaftlichen Bereichen führt schlagwort-basierte Suche nach relevanten und verwandten Publikationen zu guten Ergebnissen. Eine Literaturrecherche in Mathematik, Natur- oder Ingenieurswissenschaften bedarf jedoch, aufgrund der hohen Dichte an mathematischen Ausdrücken, spezieller zusätzlicher Strategien (Schubotz, 2017a). Einen ersten Schritt stellt dabei die Aufbereitung der mathematischen Ausdrücke dar, weil diese in der Regel für die Darstellung optimiert und als Bilder oder in speziellen Formaten wie LaTeX abgespeichert sind. Neben der Aufbereitung als maschinenlesbare Informationen, ist die automatische Erkennung von Ähnlichkeit und Verwandtheit mathematischer Ausdrücken ein wichtiger Schritt. Dies wiederum bildet die Grundlage für Such- und Empfehlungsdienste für wissenschaftliche Publikationen sowie für Verlinkungen zu ähnlichen Artikeln und Informationsquellen. Davon profitieren nicht zuletzt Schülerinnen und Schüler, sowie Studierende in naturwissenschaftlichen Fächern, die sich mathematisches Wissen mit verschiedenen digitalen Medien, wie Browsern, eBook Readern und Sprachassistenzsystemen aneignen.

Kohlhase et al. (2017) zeigen in einer Eyetracking-Studie, dass Mathematiker bei der Betrachtung mathematischer Ausdrücke mit ihren Augenbewegungen der semantischen Struktur folgen und sich die Augenbewegungen somit von denen beim Lesen natürlichsprachlicher Texte unterscheiden. Um die semantische Struktur mathematischer Ausdrücke automatisch zu erkennen ist eine semantische Anreicherung erforderlich, beispielsweise durch bedeutungstragende Informationen aus dem umgebenden Text (Schubotz et al., 2017b). Wie in Abbildung 1 gezeigt, können textuelle Informationen in Kombination mit einer Wissendatenbank dabei helfen, die dem menschlichen Experten verfügbaren Informationen, wie beispielsweise der Strukturbaum oder die Bedeutung der Variablen, für Maschinen lesbar zu

machen. Semantisch angereicherte, maschinenlesbare Mathematik ermöglicht dann die automatische Vorverarbeitung der Inhalte und eine auf den Konsumenten abgestimmte Darstellung. Dies kann neben der Berücksichtigung von speziellen Bedürfnissen wie beispielsweise eingeschränkte Sehfähigkeit oder fehlende Grundlagenkenntnisse auch eine (Such-)kontextbezogene Darstellung sein. Ginev et al. (2016) entwickeln eine Ontologie zur Speicherung von mathematischen Inhalte, welche die semantische Verknüpfung der Inhalte in verschieden strikter Formalität ermöglicht. Auf der Plattform mathhub.info können erste Ergebnisse eingesehen und eigene Inhalte erstellt werden (Iancu et al., 2014).

Grundlagenforschung als Beitrag zur Wikimedia Vision

Im Gegensatz zu mathematisch spezialisierten Projekten wie mathhub, hat Wikimedia die Vision bis 2030 „das Fundament im Ökosystem des freien Wissens“ (https://meta.wikimedia.org/w/index.php?oldid=17397593) zu sein und Informationen für jeden, unabhängig von Sprache, Herkunft und gegebenenfalls bestehenden Einschränkungen, verfügbar zu machen. Bezüglich der Weiterentwicklung des freien Wissens ergänzen sich die Forschung und die Wikimedia Stiftung (vgl. Tab. 1). Die Forschung leistet sowohl durch fachliche Kompetenzen in den einzelnen Bereichen, als auch durch konkrete Produkte ihren Beitrag zum freiem Wissen. Auf der anderen Seite stellt die Wikimedia Stiftung Infrastruktur und Daten für die Forschung bereit, die für viele Anwendungsfelder genutzt weren können (Dahm et al., 2017). Bezogen auf mathematisches Wissen gehört zur Verbesserung des Zugangs zu freiem Wissen, spezielle Wissensbestandteile wie beispielsweise mathematische Formeln, für weniger erfahrene und in diesem Fachbereich vorgebildete Leser möglichst gut aufzubereiten und zu verknüpfen. Seit 2013 existiert eine Zusammenarbeit im Bereich Mathematical Information Retrieval mit der Wikimedia Stiftung. In Wikipedia werden mathematische Formeln seit 2003 verwendet, bis 2016 wurden diese jedoch nur in Form von Pixelgrafiken dargestellt. Die Kooperation implementierte 2016 ein neues Verfahren. Seitdem liefert Wikipedia Formeln im MathML-Format aus. Dies verbessert die Möglichkeiten zur Suche, Darstellung und Weiterverarbeitung von Formeln und bildet die Grundlage für barrierefreies Fomelwissen (Schubotz & Sexton, 2016). Dabei wurde von der Forschern das Spezialwissen zur mathematischen Formeln beigesteuert. Die Softwareentwickler der Wikimedia Stiftung stellten die Einhaltung des für das Wikimedia Softwareprojekt etablierten Entwicklungsprozesses sicher. Dies vereinfacht die Wartung und erleichtert der Wikimedia Stiftung die Weiterentwicklung und Pflege des Codes. Darüber hinaus stellte die Wikimedia Stiftung ein Testcluster für MathIR Forschung bereit, mit dem auch Features

Tab. 1: Beiträge von Forschung und Wikimedia Stiftung zur Verbesserung des Zugriffs auf mathematisches Wissen

Wikimedia	**Forschung**
Infrastruktur (Server, Wartung, Softwareentwicklungsprozess)	Kompetenzen (Datenanalyse, Domänenwissen, Evaluierungsmethoden)
Daten (Nutzerstatistiken, Texte, strukturierte Daten, Multimediadaten)	Produkte (Software, Texte, Primärdaten)

getestet werden, die sich nur bedingt für den Einsatz auf Wikipedia eignen (Cohl et al., 2014). Unabhängig davon wurden die Wikipedia Daten für folgende Anwendungen genutzt (Schubotz, 2017a):

- Artikel in Deutsch, Russisch und English zum Training sowie zur Evaluierung von MathIR Methoden
- Seitenaufrufstatistiken für die Auswertung von alternativen Methoden zur Empfehlung von verwandten Artikeln
- Relationale Daten aus der Datenbank Wikidata als semantische Grundlange für die Disambiguierung mathematischer Konzepte

Die daraus entstandenen Forschungspublikationen haben den Fundus an frei zugänglichen Quellen erweitert. Zudem wird mindestens eine Publikation als Quelle in einem Wikipedia-Artikel verwendet.

Zukünftige Forschungsprojekte

Im Rahmen des DFG-geförderten Projekts "Methoden und Werkzeuge zur Verbesserung des Zugriffs auf mathematisches Wissen in digitalen Bibliotheken für Such-, Empfehlungs- und Assistenzsysteme" (GI 1259/1), welches im August 2018 unter der Leitung von Bela Gipp an der Universität Konstanz startet, ist die Verbesserung und Erweiterung der semantischen Anreicherung mathematischer Formeln geplant. Das Projekt erfolgt in enger Zusammenarbeit mit der Arbeitsgruppe von Akiko Aizawa am National Institute of Informatics in Tokio und Abdou Youssef an der University of Washingtion und dem National Institute of Informatics in den USA. Die Projektergebnisse werden auch an zum Teil experimentellen Features in Wikipedia demonstriert werden. Als erster Schritt ist geplant Methoden der künstlichen Intelligenz zu verwenden, um Autoren bei der Bearbeitung und Erstellung von mathematischen Ausdrücken zu unterstützen. Dabei werden die Methoden basierend auf dem Feedback der Autoren verbessert. Erst wenn von Menschen verifizierte semantische Informationen hoher Güte vorliegen, wird damit begonnen diese zur Verbesserung des Leseerlebnisses zu verwenden. Neben der traditionellen artikel-basierten Darstellung beschrei-

ben Corneli & Schubotz (2017) die Vision eines sprachunabhängigen, maschinenlesbaren Wiki-Projekts speziell für Mathematik. Partizipationsmöglichkeiten und neuste Updates werden im Wikidata Projekt Mathematik geteilt. Siehe http://wikidata.org/w/index.php?oldid=642233621

Danksagung: Unser Dank gilt Andrea Kohlhase und Michael Kohlhase sowie Wolfgang Sperber für die wertvollen Diskussionen sowie die Einladung zum Symposium. Weiterhin danken wir dem Projektleiter des DFG Antrags GI 1259/1 Bela Gipp, sowie dem Deutschen Akademischen Austauschdienst für die finanzielle Unterstützung.

Literatur

Cohl, H. S., McClain M. A., Sauders, B. V., Schubotz, M., Williams, J. C. (2014). Digital repository of mathematical formulae. In S. Watt et al. (Hrsg.), *Intelligent Computer Mathematics – International Conference*, (S. 419-422). Berlin: Springer

Corneli, J., Schubotz, M. (2017). math.wikipedia.org: A vision for a collaborative semi-formal, language independent math (s) encyclopedia. In Hales, C. et al. (Hrsg.). *2nd Conference on Artificial Intelligence and Theorem Proving*. http://aitp-conference.org/2017/

Dahm, E., Schubotz, M., Meuschke, N., Gipp, B. (2017). A Vision for Performing Social and Economic Data Analysis using Wikipedia's Edit History, In Barrett, R. (Hrsg.) *International World Wide Web Conference.* (S. 1627-1634). ACM Press: New York

Geuvers, H., England, M., Hasan, O., Raabe, F. Teschke, O. (Hrsg.) (2017). *Intelligent Computer Mathematics: 10th International Conference.* Berlin: Springer

Ginev, D., Iancu, M., Jucovshi, C., Kohlhase, A., Kohlhase, M., Oripov, A., Wiesing, T. (2016). The SMGloM Project and System: Towards a Terminology and Ontology for Mathematics. In Greuel, G.-M. et al. (Hrsg.) *International Conference on Mathematical Software* (S. 451-457). Berlin: Springer

Iancu, M., Jucovschi, C., Kohlhase, M., Wiesing, T. (2014). System Description: MathHub.info. In S. Watt et al. (Hrsg.), *Proc. Intelligent Computer Mathematics – International Conference*, (S. 431-434). Berlin: Springer

Kohlhase, A., Kohlhase, M., Fürsich, M. (2017). Visual Structure in Mathematical Expressions. In Geuvers 2017 (S. 208-223)

Schubotz, M., Sexton, A. P. (2016). A Smooth Transition to Modern mathoid-based Math Rendering in Wikipedia with Automatic Visual Regression Testing. In Kohlhase, M. et al (Hrsg.). *Intelligent Computer Mathematics: 9th International Conference* (S. 132-145) Berlin: Springer

Schubotz, M. (2017a). Augmenting Mathematical Formulae for More Effective Querying & Efficient Presentation. Berlin: Epubli

Schubotz, M., Krämer, L., Meuschke, N., Hamborg, F., Gipp, B. (2017b). Evaluating and Improving the Extraction of Mathematical Identifier Definitions. In Gareth, J. (Hrsg.) *International Conference of the Cross-Language Evaluation Forum for Europ. Languages* (S. 82-94). Berlin: Springer

Schubotz, M., Meuschke, N., Hepp, T., Cohl, H. S., Gipp, B. (2017c). VMEXT: A Visualization Tool for Mathematical Expression Trees. In Geuvers (2017) (S. 340-355)