Leibniz-Gemeinschaft

Abschließender Sachbericht

MathSearch
-Analyse und Suche in mathematischen Formeln-

Leibniz-Einrichtung: FIZ Karsruhe Leibniz-Institut für Informationsinfrastruktur
Aktenzeichen: SAW-2011-FIZ_KA-2
Projektlaufzeit: 01.01.2012 – 30.11.2015
Ansprechpartner: Dr. Wolfram Sperber

Abschlußbericht für das SAW-Projekt MathSearch

– Analyse und Suche in mathematischen Formeln –

FIZ Karlsruhe
Leibniz-Institut für Informationsinfrastruktur

FIZ Karlsruhe
Leibniz-Institut für Informationsinfrastruktur

JACOBS
UNIVERSITY

Aktenzeichen	SAW-2012-FIZ_KA-2
Projektzeitraum	1.1.2012 - 30.11.2015
Antragssteller	FIZ-Karlsruhe Leibniz-Institut für Informationsinfrastruktur, Hermann-von Helmholtz-Platz 1 76344 Eggenstein-Leopoldshafen

Wolfram Sperber
FIZ Karlsruhe
Leibniz-Insitut für Informationsinfrastruktur, zbMATH
Franklinsstr.11, D-10587 Berlin

Michael Kohlhase
Jacobs University Bremen

Campus Ring 1, D-28759 Bremen

FIZ Karlsruhe
Leibniz-Institut für Informationsinfrastruktur

JACOBS UNIVERSITY

Inhaltsverzeichnis

1 Ziele; Inhalte und wesentliche Ergebnisse

Das MathSearch Projekt wurde innerhalb des SAW Verfahrens in der Förderlinie „Risikoreiche Forschung“ 2011-2015 durchgefürt. Ziel des Vorhabens war es, neue Methoden für die mathematische Formelsuche zu entwickeln und am Beispiel der Datenbank zbMATH zu etablieren. Dazu sollten die folgenden Probleme gelöst werden:

1. Extraktion und Standardisierung der mathematischen Formeln der Datenbank zbMATH,
2. Semantische Auswertung der vorhandenen mathematischen Formeln und deren Kontext,
3. Konvertierung der Formeln nach Content MathML und Aufbau eines Formelindexes,
4. Entwicklung einer Suchmaschine, die auch eine Formelsuche bietet.

Im Rahmen des Projektes wurden innovative Methoden und Werkzeuge für die Untersuchung und Indexierung mathematischer Formeln entwickelt und implementiert. Die wichtigsten Projektergebnisse im Überblick:

- Im Projekt wurden Konzepte für die Verarbeitung mathematischer Formeln entwickelt und prototypisch erprobt. Dazu wurden die TEX-kodierten Formeln der Datenbank zbMATH nach MathML konvertiert und im MathWebSearch-System indexiert (Problem 1).
- Eine Struktur-Formelsuche wurde in die Datenbank zbMATH als zusätzliche Suchfunktionalität integriert und steht seit 2014 allen Nutzern der Datenbank zbMATH zur Verfügung (Problem 4). Dazu wurden die Formeln nach Content MathML transformiert (Probleme 3). Allerdings führt die Konvertierung häufig zu Termen, die verschiedene inhaltliche Bedeutung haben können, etwa ob P ein Polynom oder die Wahrscheinlichkeit bezeichnet. Eine semantische Disambiguierung kann mittels einer semantischen TeX-Kodierung und/oder Kontextanalyse erfolgen. Mittels der TEX-Erweiterung STEX gibt es ein Konzept und ein Werkzeug für eine eindeutige semantische Kodierung mathematische Formeln. Bisher gibt es aber noch keinen Standard für die STEX Kodierung mathematischer Symbole.
- Für die Erweiterung zu einer semantischen Formelsuche wurde das Konzept eines semantischen mathematischen Glossars entwickelt und implementiert. Dieser enthält u.a. auch eine semantische Kodierung in STEX wichtiger mathematischer Terme (Problem 3).

Die Projektaktivitäten und Projektergebnisse werden im Folgenden detailliert vorgestellt.

2 Das Umfeld für die Formelsuche

Mathematische Symbole und Formeln spielen in der Mathematik eine besondere Rolle, da sie anders als die normale Sprache in der Lage sind, komplexe Zusammenhänge in der Mathematik in komprimierter Form und wesentlich exakter als mit der normalen Umgangssprache möglich darzustellen. Allerdings stellen die Darstellung mathematischer

Symbole und Formeln, deren Erschließung und die Suche aus verschiedenen Gründen eine besondere Herausforderung dar:

- Viele mathematische Symbole und Formeln (etwa Brüche) haben eine zweidimensionale Darstellung.
- Es gibt viele spezielle mathematische Symbole. Dafür werden Zeichen aus verschiedenen Alphabeten als mathematische Symbole verwendet.

Diese Herausforderungen haben zur Entwicklung spezieller Auszeichungssprachen für die Mathematik geführt, etwa TEX und MathML. TEX hat sich als flexibles and elegantes Input Format für die Mathematik durchgesetzt und bietet neben dem Layout auch semantische Informationen zu mathematischen Symbolen und Formeln, allerdings nur im geringem Umfang. Durch zusätzliche semantische Annotationen kann aber das TEX Format angereichert werden.

Die Mathematical Markup Language, MathML, kann sowohl zur Darstellung (Presentation MathML) als auch zur semantischen Darstellung (Content MathML) und zur Darstellung mathematischer Symbole und Formeln benutzt werden. OMDoc bietet Metastrukturelemente für Dokumente mathematischen Inhalts und erlaubt ein semantisches Markup dieser Inhalte, etwa mathematischer Theorien oder wissenschaftlicher Artikel. Da die Eingabedaten typischerweise nur über wenige semantische Informationen verfügen, führt die Konvertierung i.A. zu verschieden semantisch interpretierbaren Ergebnissen.

Eine einfache Textsuche ist für die Mathematik nicht ausreichend. Mittels Textsuche kann nicht nach Symbolen und Formeln gesucht werden. Auch eine TEX-Suche ist nur bedingt für die Formelsuche geeignet, da die TEX-Kodierungen der Symbole und Formeln nur rudimentäre semantische Informationen enthalten. Neben der semantischen Ambiguität der Symbole und Formeln kommt hinzu, dass

- Variablen in den Symbolen und Formeln enthalten sein können, die für die Semantik unwesentlich sind
- das TEX-Format viele äquivalente Formatierungen zulässt (die sich etwa bei der Klammersetzung oder durch Leerzeichen unterscheiden können). Eine syntaktische Normalisierung ist in TEX wegen der Macros in der Praxis unmöglich.

Um das Ziel des MathSearch Projektes, die Entwicklung eines langfristig tragfähigen Konzepts für die mathematische Formelsuche und dessen Implementierung für die Datenbank zbMATH zu erreichen, wurde im Projekt die Konvertierung nach Content MathML und die Untersuchung der daraus erzeugten Graphenstruktur als Ansatz für die Formelsuche verwendet.

3 Projektverlauf und Projektergebnisse

3.1 Analyse der zbMATH Daten

Die zbMATH Daten waren das Testbed, um die an der JUB entwickelten Analyseverfahren und Suchstrategien für mathematische Formeln zu erproben.

Zu Beginn des Projekts erfolgte eine Analyse der Daten der Datenbank zbMATH hinsichtlich der Darstellung mathematischer Zeichen und Formeln. Die in zbMATH enthaltenen Formeln sind in TEX kodiert. TEX erlaubt eine sichere Identifizierung mathematischer Entitäten in einem Text. Zunächst wurde versucht (analog zu den Bag-of-Words Ansätzen in der Textanalyse) mit Häufigkeitsanalysen für mathematische Symbole und

Formeln zu arbeiten. Dazu wurden die ca. 10.000.000 TEX kodierten mathematischen Symbole und Formeln in zbMATH untersucht. Es wurde aber festgestellt, dass die TeX-Kodierungen bei inhaltlich identischen längeren Formeln sich i.A. unterscheiden, dass also Häufigkeiten der TEX-kodierten Ausdrücke nicht zur Analyse verwendet werden können. Die Ursachen dafür sind – wie oben kurz dargestellt – verschiedenartig:

- die mathematischen Schreibweise für ein- und denselben mathematischen Sachverhalt sind nicht eindeutig (z.B. können Differentialgleichungen in verschiedener Form dargestellt werden)
- mathematische Formeln können Variable enthalten, die durch andere Variablen ersetzt werden können, ohne dass sich die Bedeutung ändert. Diese Variablen sind in den zbMATH Daten allerdings nicht gekennzeichnet. Inhaltlich identische Formeln können damit nicht als solche erkannt werden.
- die TEX Kodierung ist nicht eindeutig, das heißt, unterschiedliche TEX-Kodierungen haben dieselbe Darstellung (z.B. durch unterschiedliche Verwendung von TEX-Gruppen oder Macros, diese können nutzerdefiniert sein).
- die TEX Kodierung der zbMATH Daten war zu Projektbeginn nicht einheitlich, was im Wesentlichen historische Ursachen hat: Die zbMATH Daten werden seit mehr als 30 Jahren im TEX erstellt, das TEX Format hat sich aber in dieser Zeit verändert.

Mit seinen Makropakten ist heute LATEX zum de-facto Standard für TEX geworden. Die zbMATH Daten machten daher eine Vorbehandlung (Konvertierung nach LATEX) erforderlich, die gemeinsam von FIZ und JUB vorgenommen wurde. Dazu wurde ein Konzept für den Datenaustausch zwischen FIZ Karlsruhe und JUB entwickelt und implementiert.

3.2 Weitere Aufbereitung der Daten: Konvertierung nach MathML: LATEXML

Bedeutend besser als das TEX- (bzw. das LATEX-) Format eignet sich das MathML-Format für die Suche:

- Ein wesentlicher Vorteil von MathML gegenüber TEX ist, dass es sich bei MathML um eine XML-Sprache ohne Macrofunktionalität (der Nutzer kann also keinen eigenen Programmcode einfügen und ausführen) handelt. Insbesondere sind mathematische Formeln als statische Bäume repräsentiert, was die Analyse wesentlich vereinfacht. Damit lassen sich die Werkzeuge der Graphentheorie zur Analyse und zum Vergleich von Formeln und Formelteilen einsetzen, etwa um Differentialgleichungen oder deren charakteristische Bestandteile (z.B. Differentialoperatoren) zu vergleichen.
- Die XML Darstellungen der Formeln lassen sich durch die Strukturanalysen teilweise einfacher normalisieren: So können etwa überflüssige Klammern erkannt und entfernt werden. Verknüpfungen zwischen verschiedenen Symbolen können disambiguiert werden.

MathML-Darstellungen umfassen Presentation MathML und Content MathML. Während Presentation MathML auf die Darstellung mathematischer Symbole und Formeln im Web abzielt und wie TEX nur in geringem Umfang semantische Aussagen enthält, ist der Fokus von Content MathML die semantische Darstellung der mathematischen Symbole und Formeln. Content MathML ist daher für die Formelsuche weitaus interessanter als Presentation MathML. In den letzten Jahren wurde eine Reihe von Tools für die Konvertierung des TEX-Codes nach Presentation MathML entwickelt. Der derzeit am weitesten

entwickelte Konverter ist LaTeXML, der nicht nur eine Konvertierung nach Presentation MathML sondern – so weit wie möglich – auch nach Content MathML vornimmt. LaTeXML ist ein von Bruce Miller (NIST) und Deyan Ginev (JUB) entwickeltes Konzept und Software Tool, das TeX auf Presentation MathML und Content MathML abbildet, insbesondere kann LaTeXML zusätzliche semantische Annotationen nach Content MathML transformieren. Im Rahmen des MathSearch Projektes wurde an der Weiterentwicklung des Konverters LaTeXML gearbeitet, insbesondere an der Konvertierung nach Content MathML. Die Konvertierung der TeX-kodierten Symbole und Formeln kann aber – ohne zusätzliche Informationen – nur zu einer rudimentären Semantifizierung der mathematischen Symbole und Formeln führen. Symbole und Formeln, die häufig verschiedene mathematische Bedeutungen haben können, müssen durch weitere Untersuchungen disambiguiert werden. Das kann etwa durch eine Analyse des Kontexts eines Symbols oder einer Formel erfolgen oder durch eine Semantifizierung des TeX Codes, etwa mittels des Makropakets STeX, das die semantische Annotation von Symbolen und Formeln erlaubt.

Im Projekt wurde von der JUB ein Build-System entwickelt, das den gesamten Konvertierungsprozess abwickelt und kontrolliert. Die Input-Daten, eine LaTeX formatierte Datei, werden mittels LaTeXML nach MathML konvertiert, LaTeX Fehler als auch Probleme bei der Konvertierung werden registriert, ausgewertet und angezeigt. Ohne ein solches Buildsystem sind Textkorpora mit mehreren Millionen Dokumenten und Fehler/Statusberichte des Konverters nicht beherrschbar.

Für die LaTeXML-Konvertierung wurde ein Werkzeug entwickelt, das grosse Mengen von Dokumenten ohne Neustart verarbeiten kann. Durch den Wegfall des Startup-Overheads sinken die Verarbeitungszeiten von ca 30 Sekunden pro Dokument auf Sekundenbruchteile und die Konvertierung des gesamten zbMATH Korpus von Prozessorjahren auf Prozessortage.

Die mathematischen Symbole und Formeln der Datenbank zbMATH wurden indexiert und ein separater Formelindex aufgebaut. Für die Formelsuche wurde ein Interface entwickelt und in die Suchoberfläche der Datenbank zbMATH integriert. Die Formelsuche wurde zum ICM (Internationalen Congress of Mathematicians) August 2014 in Seoul der mathematischen Öffentlichkeit vorgestellt. Mit der implementierten Suche in zbMATH wurde ein erster wichtiger Schritt in Richtung einer leistungsfähigen Formelsuche gemacht. Er erweitert entscheidend die Retrievalmöglichkeiten in mathematischen Dokumenten: Die Formelsuche erlaubt eine präzise Suche nach mathematischen Objekten und Konzepten auf Dokumentebene und ist unverzichtbar für den geplanten Aufbau mathematischer Bibliotheken. Die Nutzung der verwendeten Web Standards garantiert die nachhaltige Nutzbarkeit des Konzepts für die Behandlung mathematischer Symbole und Formeln.

Die Formelsuche erfordert eigentlich eine MathML-kodierte Anfrage, was aber mit hohen Aufwand und deshalb nicht nutzerfreundlich ist, daher erlauben wir die Kodierung der Anfrage in LaTeX, wo Anfragevariablen - diese stehen für beliebige Teilformeln - mit ? gekennzeichnet werden (siehe Abbildung 1) und konvertieren sie mit dem LaTeXML-Webservice nach Content MathML. Die Anfrage kann aus einem einzelnen mathematischen Symbol, einer kompletten Formel oder einem beliebigen Bestandteil einer mathematischen Formel bestehen. Zur Kontrolle der Korrektheit des Anfrage wird zusätzlich die Web Präsentation der Formel angezeigt.

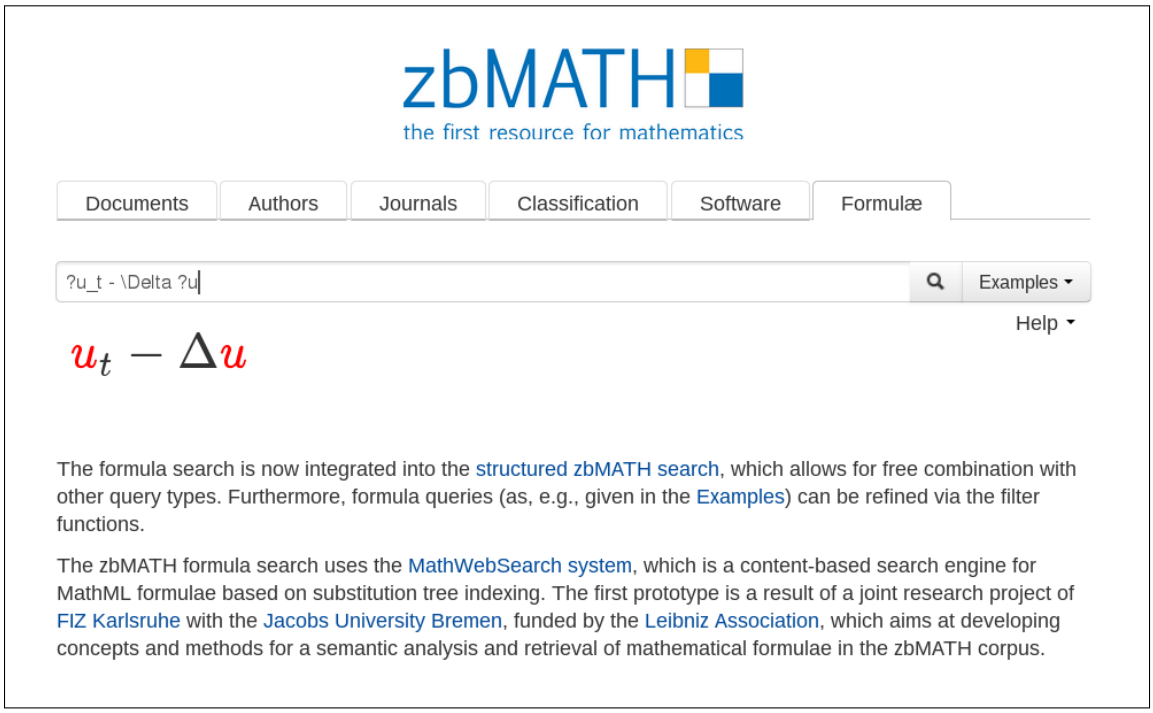

Abbildung 1: Screenshot der Formelsuche in zbMATH (hier: Suche nach der Formel $u_t - \Delta u$)

Die Bearbeitung der Anfrage erfolgt analog zur Bearbeitung der zbMATH-Daten. Die Anfrage wird zunächst nach MathML konvertiert und es wird danach im Formelindex nach Termen gesucht, die eine identische Baumstruktur zu der gesuchten Formel enthalten.

Die Formelsuche ist mit anderen Feldsuchen beliebig kombinierbar, ist also eine weitere Facette für das Retrieval in den zbMATH Daten. Abbildung 2 zeigt einen Ausschnitt der Trefferliste für die Anfrage in Abbildung 1. Der Suchterm, der aus dem Bereich der partiellen Differentialgleichungen stammt, wurde in insgesamt 476 Dokumenten gefunden.

Die Formelsuche arbeitet stabil und effizient, wird allerdings bisher nur in geringem Umfang benutzt. Das hat verschiedene Gründe:

- Ein wesentlicher Grund ist, dass Formelsuche eine neue Retrieval Funktionalität darstellt und anspruchsvoller als eine Textsuche ist. Insbesondere wird eine TeX-Kodierung der Formel vorausgesetzt, zudem müssen Variablen besonders gekennzeichnet werden Falls mehrere Darstellungen eines Symbols oder einer Formel existieren, wie z.B. für gewöhnliche und partielle Ableitungen, müssen derzeit alle Darstellungen einzeln abgefragt werden. Im Projekt wurden Untersuchungen zur Web Interfaces für mathematische Informationsdienste durchgeführt [**Koh14a**], die wichtige Hinweise für die Weiterentwicklung der Suchoberflächen erbracht haben: neue Funktionalitäten in den User Interfaces setzen sich nicht von allein durch, Mathematiker bevorzugen minimale User Interfaces und erwarten eine hohe Relevanz der Ergebnisse.
- Ein weiterer Grund ist darin zu sehen, dass Formeln in den zbMATH Daten eine untergeordnete Rolle spielen. Die zbMATH Daten, insbesondere Abstract oder Re-

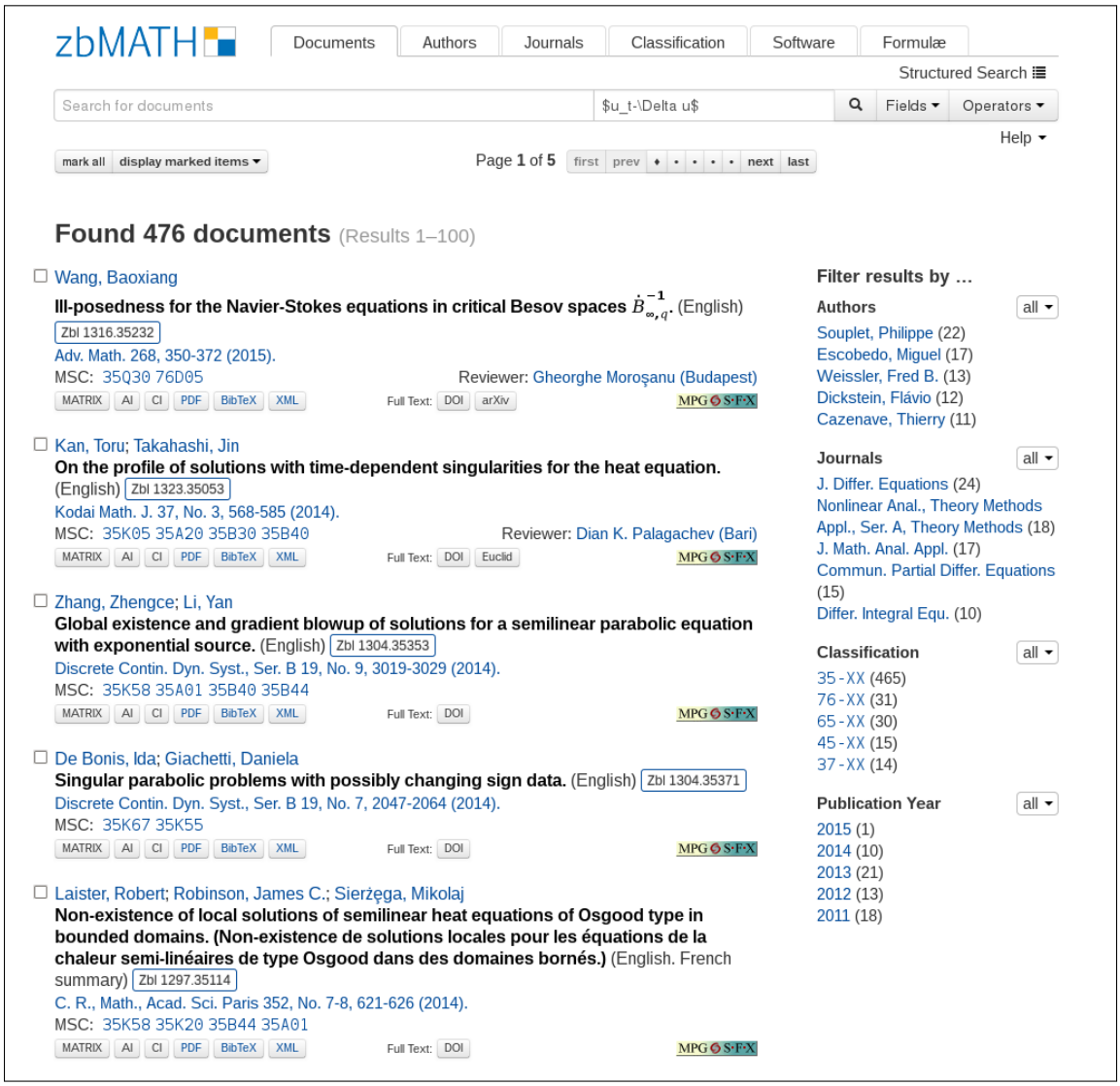

Abbildung 2: Screenshot eines Treffers der Anfrage aus Bild 1

view, Titel und Keyword weisen im Vergleich zu Volltexten nur relativ wenige Formeln auf. Das hängt einerseits vom mathematischen Gebiet ab. Generell sind aber in den Volltexten der Publikationen mehr Formeln als in den Abstracts oder Reviews zu erwarten (etwa durch die präzise Formulierung der Sätze und der Beweise, die in den Abstracts oder Reviews nur angesprochen werden).

- Zudem fehlt bisher eine Verlinkung der Namen und mathematischen Symbolen oder Formeln. Es ist also keine integrierte Suche nach Texten und Formeln möglich. Text- und Formelindex sind unabhängig (können aber kombiniert werden). Eine Suche etwa nach dem Suchwort „Laplace equation“ findet nur dann die entsprechende Formel, wenn die Phrase „Laplace equation“ auch im Text vorkommt.
- Der fehlende semantische Bezug der TEX Daten kann dazu führen, dass bei der Formelsuche auch solche gefunden werden, die verschiedenen mathematischen Gebieten entstammen (also nicht zu der vom Nutzer intendierten Frage passen). An-

ders ausgedrückt, die bisher realisierte Formelsuche ist noch nicht mathematisch-semantisch, sondern stellt eher eine erweiterte formelsyntaktische Suche dar.

4 Semantische Formelsuche über Mathematische Terminologie

Die Weiterentwicklung der Formelsuche bedarf neuer innovativer Konzepte. Das Problem liegt im Wesentlichen darin, dass die TEX-kodierten Symbol- und Formeldarstellungen zu wenig semantische Information enthalten. Die Anfragen der Nutzer sind aber in der Regel inhaltlicher Natur.

Eine vielversprechende Möglichkeit ist der Aufbau einer semantischen Terminologie für die Mathematik und ihrer Anwendungsgebiete. Im Projekt wurde ein innovatives Konzept für eine solche Terminologie, das „Semantic Multilingual Glossary on Mathematics“ (SMGloM), entwickelt.

4.1 Das Konzept des Semantic Multilingual Glossary on Mathematics

Enzyklopädien und Glossare sind wichtige Wissensspeicher und Referenzquellen. Die existierenden mathematischen Enzyklopädien wie Wikipedia oder Encyclopedia of Mathematics, sind umfangreiche und relevante mathematische Vokabulare mathematischer Objekte und Konzepte, die aber nur TEX-kodierte Formeln und keine semantischen Annotationen enthalten. Synonyme und homonyme mathematische Symbole und Schreibweisen mathematischer Terme können in den existierenden Glossaren somit nicht identifiziert werden bzw. können nicht maschinell verarbeitet werden.

Das Ziel des SMGloM Aktivität besteht im Aufbau einer multilingualen Ontologie der Mathematik, die die mathematischen Objekte und Konzepte in ihrem mathematischen Kontext darstellt und eindeutige semantische Zuordnungen einer Definition zu den dafür gebräuchlichen Begriffen und den verwendeten Symbolen und Formeln einführt. Dazu werden die semantischen Relationen zwischen den SMGloM Termen explizit dargestellt und Identifier für die SMGloM Terme und den Symbole und Formeln eingeführt. SMGloM ist das erste Vokabular in der Mathematik, das die Beziehungen zwischen mathematischen Objekten oder Konzepten darstellt, und für jedes mathematische Objekt oder Konzept die dafür verwendeten Symbolen oder Formeln und deren Bezeichnungen definiert und maschinell verarbeitbar macht.

Die Relationen zwischen den Begriffen umfassen die Abhängigkeiten eines SMGloM Terms von anderen SMGloM Termen, also der Terme, die zur Definition des Terms verwendet werden. Zudem erlaubt SMGloM auch die Modellierung anderer Abhängigkeiten und Beziehungen zwischen den Termen. z.B. von Äquivalenzen von SMGloM Termen (dann, wenn ein Term auf unterschiedliche Art und Weise definiert werden kann). Die Relationen zwischen den SMGloM Termen lassen sich als Graphen darstellen und erlauben den Nutzern einen intuitiven und semantischen Zugang zu einem gesuchten mathematischen Term und seinem Umfeld.

4.2 Das Datenmodell von SMGloM

Das Datenmodell von SMGloM basiert auf dem OMDoc/MMT-Konzept, das von M. Kohlhase und seiner Arbeitsgruppe entwickelt worden ist. Es erlaubt sowohl eine seman-

tische Darstellung der Bestandteile der SMGlom Terme als auch deren Relationen zu anderen Termen. SMGloM bettet die mathematischen Terme (Objekte und Konzepte) in eine konzeptionelle Graphenstruktur ein und umfasst

- einen formalen Identifier für den mathematischen Term (mathematisches Objekt/Konzept)
- eine (multilinguale) Definition des mathematischen Terms
- eine Darstellung der Abhängigkeiten eines Terms zu anderen Termen (Benennung der Glossarterme, die zur Definition des Terms verwendet werden)
- die verschiedenen Bezeichnungen des mathematischen Terms
- die für den Term verwendeten mathematischen Symbole

Die Glossareinträge werden in einem Modul zusammengefasst, der aus $n + 1$ Dateien besteht, einer sprachunabhängigen Datei, die den Identifier, die Abhängigkeiten und die Symbole enthält sowie n sprachabhängigen Dateien („Language Bindings“), die für jede Sprache die Definition und die verschiedenen Bezeichnungen des Terms enthalten.

Mathematische Terme können oft auf mehr als eine Art definiert werden. Äquivalenzen zwischen verschiedenen Definitionen beruhen oft auf tiefliegenden mathematischen Zusammenhängen. Das SMGLOM Konzept sieht vor, dass für jede Definition ein eigener SMGloM Eintrag erstellt wird. Nicht-hierarchische Relationen zwischen mathematischen Objekten und Konzepten können in SMGloM durch sogenannte „Views“ zwischen verschiedenen SMGloM Einträgen modelliert werden, insbesondere auch Äquivalenzen von SMGloM Einträgen, also Terme, die durch Definitionen gegeben sind, die zueinander äquivalent sind.

4.3 Entwicklungsstand

Der im Projekt entwickelte Prototyp von SMGloM umfasst derzeit etwa 500 mathematische Module, die 1.500 verschiedensprachige Definitionen enthalten. Die Einträge entstammen unterschiedlichen mathematischen Gebieten. Die Einträge wurden im STEX-Format von Mitarbeitern der JUB und FIZ manuell erstellt, was mit einem hohen Aufwand verbunden war. Die mit der Erstellung der Einträge gemachten Erfahrungen wurden unmittelbar zur Anpassung und Verfeinerung des SMGloM Modells eingesetzt.

Das Konzept von SMGloM bietet für die Weiterentwicklung der Formelsuche entscheidende Vorteile. Wie oben dargestellt, führt die Bearbeitung der Symbole und Formeln zu Baumstrukturen, die verschiedene inhaltliche Bedeutung haben können. Für die Disambiguierung der möglichen inhaltlichen Interpretationen lässt sich SMGloM wie folgt einsetzen:

- Ermittlung aller möglichen SMGloM Terme für ein Symbol, Formel oder Teilformel (Welche SMGloM Einträge haben dieselbe Darstellung wie das untersuchte Symbol oder die Formel?). Damit werden die möglichen mathematischen Objekte oder Konzepte identifiziert.
- Überprüfung der in einer Formel auftretenden mathematischen Terme auf deren Kompatibilität zueinander und Abgleich mit Textinformationen und den anderen Formeln in einem mathematischen Text.

Bemerkungen:

- Die Kontextanalyse kann um andere Kriterien erweitert werden. So soll beispielsweise die MSC Klassifizierung der untersuchten Dokumente berücksichtigt werden. Das lässt eine heuristische Disambiguierung der Symbole und Formeln zu, etwa ob

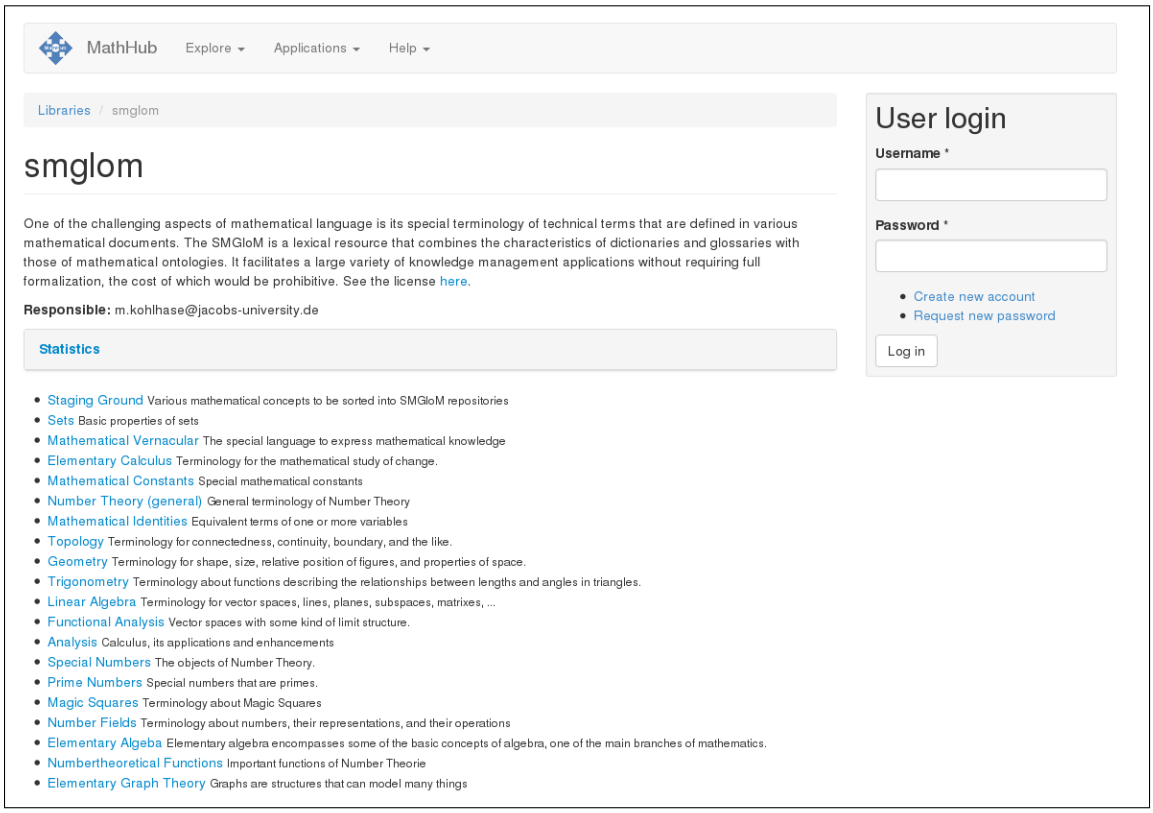

Abbildung 3: Screenshot des Homepage von SMGloM

es sich bei „P“ um ein Symbol für ein Polynom oder eher um das Symbol der Wahrscheinlichkeit handelt.

- SMGloM kann weitere nützliche Aufgaben übernehmen, z.B. die eines normalen Glossars, zur Übersetzung mathematischer Begriffe (so weit die entsprechenden Language Bindings vorhanden sind), oder zur Semantifizierung und Darstellung eigener mathematischer Dokumente - siehe Abbildung 4.

Die Überlegungen, die zur Entwicklung von SMGloM geführt haben, sind sehr komplex und gehen über die Möglichkeiten der derzeitig existierenden mathematischen Glossare im Web deutlich hinaus. Durch die speziellen Anforderungen der Mathematik (Formeldarstellung, Dualität von Text und Formeln) sind die Standard Technologien des Semantic Web nicht ausreichend für SMGloM, sondern müssen für die Semantifizierung mathematischer Inhalte angepasst und erweitert werden.

4.4 Nutzeranforderungen an SMGloM

Das SMGloM-System soll in den nächsten Jahren - über das MathSearch-Projekt hinaus - weiterentwickelt werden und vor allem skaliert werden. Um allerdings für die mathematische Community akzeptabel und sogar attraktiv zu werden, muss das System einige Nutzeranforderungen erfüllen:

- **eine hohe inhaltlich Qualität jedes Eintrags**
  Dazu zählen die inhaltliche Korrektheit der Definition (die einen Nachweis der Quellen enthalten sollte), die Bezeichnungen der Terme, die zur Definition verwen-

Abbildung 4: Screenshot für die englische Definition des Begriffs Graph in SMGloM

det werden sowie die Symbole oder Formeln, die für den Term verwendet werden

- (langfristig) **eine hohe Abdeckung des mathematischen Vokabulars**
  Für die semantische Identifizierung mathematischer Symbole und Formeln ist eine möglichst umfassende Abdeckung des mathematischen Vokabulars notwendig. Dadurch werden mögliche Bedeutungen angezeigt, die dann zur Disambiguierung benutzt werden können. Langfristig bietet nur eine hohe Vollständigkeit des mathematischen Vokabulars ausreichend Potential für die Disambiguierung.
- **formale Korrektkeit der Einträge**
  Die formale Korrektheit der SMGloM-Terme (Struktur und Syntax) ist zwingend für die maschinelle Nutzung des Glossars. Das Modell von SMGloM und dessen Realisierung sind nicht einfach. Im Projekt wurden erste Werkzeuge für die maschinelle Unterstützung der Erstellung, Anzeige, und Kontrolle der strukturellen / syntaktischen Korrektheit der SMGloM Terme entwickelt.
- **einfache Referenzierbarkeit und Wiederverwendbarkeit der in SMGloM definierten Symbole, Formeln und Begriffe**
  Das Glossar stellt einen eigenen Namespace für mathematische Terme dar. Die SMGloM Terme und deren Bestandteile, etwa die Identifier der Symbole oder Formeln, können von der Community für die Semantifizierung eigener Dokumente verwendet werden.

Bemerkungen: Aus den Anforderungen ergeben sich Konsequenzen für die Weiterentwicklung von SMGloM.

- **Verteiltes System**: Um den inhaltlichen Anforderungen gerecht zu werden, bedarf es Expertenwissens. Bei der heute erreichten Spezialisierung der Mathematik ist es allerdings nicht möglich, dass eine einzelne Person oder eine kleine Gruppe die Expertise für alle mathematischen Gebiete hat. Die Erstellung des Glossar muss also verteilt erfolgen. Das Glossar setzt auf eine breite Unterstützung durch die mathematische Community.
- **Technisches Konzept**: Das System muss einen verteilten Input unterstützen und die strukturelle und syntaktische Korrektheit der Einträge sichern, was im Rahmen des bisherigen Projekte nur modelliert und prototypisch implementiert werden konnte. Dazu zählen etwa:
  - Werkzeuge zur Unterstützung der Eingabe (Ist der beabsichtigte Term in SM-

GloM bereits vorhanden?, Welche Enträge aus dem Umfeld des beabsichtigten Eintrags sind in SMGloM vorhanden?, Wie kann die strukturierte Eingabe der SMGloM Einträge unterstützt werden?, Wie kann die syntaktische Korrektheit der Einträge verifiziert werden?) Es wird für das SMGloM System - ähnlich wie für die oben beschriebene Konvertierung der zbMATH Daten nach MathML - ein Build-System benötigt, dass die Einträge auf syntaktische und strukturelle Korrektheit überprüft und den Autoren Rückmeldungen (Hinweise und Fehlermeldungen) liefert.

- Werkzeuge zur Datenpflege Änderungen, etwa Erweiterungen, der Datenstruktur von SMGloM müssen automatisch erfolgen/unterstützt werden.
- Eine Organisationsstruktur und Workflowsystem für die Erstellung und Evaluierung (Peer Reviewing) der SMGloM-Einträge muss aufgebaut werden.

Für die Entwicklung der Softwareplattform und der Werkzeuge zur Kontrolle und den Betrieb von SMGloM ist eine stabile Entwicklergruppe notwendig.

Ein erster Prototyp des SMGloM-Systems ist im Rahmen des MathHub.info Portals [Ian+14; MH] realisiert worden, und und steht im Web zur Verfügung.

5 Veröffentlichungen und Vorträge

Die Projektergebnisse wurden in den wichtigen Fachzeitschriften publiziert und auf verschiedenen Tagungen vorgestellt, die im Folgenden aufgelistet werden:

- Jahrestagungen der Deutschen Mathematiker-Vereinigung 2013 (Innsbruck), 2014 (Poznan) und 2015 (Hamburg)
- Conference on Intelligent Computer Mathematics 2014 (Coimbra), 2015 (Washington)
- NTCIR (NII Textbeds and Communications for Information Access Research) 2013 (Tokio), 2014 (Tokio)
- ECDA (European Conference on Data Analysis) (2015, Colchester)

6 Zusammenfassung

Die im Projekt erstellten Konzepte und Werkzeuge stellen einen Ausgangpunkt und wichtigen Schritt für die Entwicklung einer Methode für mathematischen Informationsretrieval dar, der die Struktur und Spezifika der Mathematik, etwa die üblichen „flexifomalen“ Darstellungen mathematischer Sachverhalte, die aus einem Mix von Text und Formeln bestehen, berücksichtigt.

Das Projekt ist – wie die obigen Ausführungen zeigen – nicht abgeschlossen, sondern bietet großes Potential für die Darstellung und Verarbeitung mathematischen Wissens im Web: Mit dem Glossar SMGloM im Projekt ein neues Werkzeug entwickelt, wodurch sich die im Projekt entwickelte Formelsuche zu einer semantischen Suche erweitern lässt. Dieses Konzept verwendet Semantic Web Methoden und adaptiert und erweitert diese für den Bereich der Mathematik.

Das im Projekt entwickelte Konzept und die Werkzeuge sind wichtige Resultate für die künftige Bereitstellung von mathematischen Informationen und zum Zugang zu diesen.

References

[Aiz+14] Akiko Aizawa et al. "NTCIR-11 Math-2 Task Overview". In: *NTCIR 11 Conference*. Ed. by Noriko Kando, Hideo Joho, and Kazuaki Kishida. Tokyo, Japan: NII, Tokyo, 2014, pp. 88–98. URL: http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings11/pdf/NTCIR/OVERVIEW/01-NTCIR11-OV-MATH-AizawaA.pdf.

[AKO13] Akiko Aizawa, Michael Kohlhase, and Iadh Ounis. "NTCIR-10 Math Pilot Task Overview". In: *NTCIR Workshop 10 Meeting*. Ed. by Noriko Kando and Kazuaki Kishida. Tokyo, Japan: NII, Tokyo, 2013, pp. 1–8. URL: http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings10/pdf/NTCIR/OVERVIEW/01-NTCIR10-OV-MATH-AizawaA.pdf.

[DK13] Jan Wilken Dörrie and Michael Kohlhase. "OpenMathMap: accessing math via interactive maps". In: *MathUI, OpenMath, PLMMS, and ThEdu Workshops and Work in Progress at the Conference on Intelligent Computer Mathematics*. (Bath, UK, July 8–12, 2013). Ed. by Christoph Lange et al. CEUR Workshop Proceedings 1010. Aachen, 2013. URL: http://ceur-ws.org/Vol-1010/paper-12.pdf.

[HK16] Radu Hambasan and Michael Kohlhase. "Faceted Search for Mathematics". In: *MACIS 2015: Sixth International Conference on Mathematical Aspects of Computer and Information Sciences*. LNAI. in press. Springer Verlag, 2016. URL: http://kwarc.info/kohlhase/papers/macis15.pdf.

[HKP14] Radu Hambasan, Michael Kohlhase, and Corneliu Prodescu. "MathWebSearch at NTCIR-11". In: *NTCIR 11 Conference*. Ed. by Noriko Kando, Hideo Joho, and Kazuaki Kishida. Tokyo, Japan: NII, Tokyo, 2014, pp. 114–119. URL: http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings11/pdf/NTCIR/Math-2/05-NTCIR11-MATH-HambasanR.pdf.

[Ian+14] Mihnea Iancu et al. "System Description: MathHub.info". In: *Intelligent Computer Mathematics 2014*. Conferences on Intelligent Computer Mathematics. (Coimbra, Portugal, July 7–11, 2014). Ed. by Stephan Watt et al. LNCS 8543. Springer, 2014, pp. 431–434. URL: http://kwarc.info/kohlhase/submit/cicm14-mathhub.pdf.

[KJK14] Noriko Kando, Hideo Joho, and Kazuaki Kishida, eds. *NTCIR Workshop 11 Meeting*. Tokyo, Japan: NII, Tokyo, 2014.

[KK13] Noriko Kando and Kazuaki Kishida, eds. *NTCIR Workshop 10 Meeting*. Tokyo, Japan: NII, Tokyo, 2013.

[Koh+13] Michael Kohlhase et al. "Zentralblatt Column: Mathematical Formula Search". In: *EMS Newsletter* (Sept. 2013), pp. 56–57. URL: http://www.ems-ph.org/journals/newsletter/pdf/2013-09-89.pdf.

[Koh14a] Andrea Kohlhase. "Design of Search Interfaces for Mathematicians". In: *MathUI, OpenMath, PLMMS, and ThEdu Workshops and Work in Progress at the Conference on Intelligent Computer Mathematics*. (Coimbra, PT, July 7–12, 2013). Ed. by Matthew England et al. CEUR Workshop Proceedings 1180. Aachen, 2014. URL: http://ceur-ws.org/Vol-1186/paper-02.pdf.

[Koh14b] Andrea Kohlhase. "Math Web Search Interfaces and the Generation Gap of Mathematicians". In: *Mathematical Software - ICMS 2014 - 4th International Congress*. Ed. by Hoon Hong and Chee Yap. Vol. 8592. LNCS. Springer, 2014, pp. 586–593. DOI: 10.1007/978-3-662-44199-2_88. URL: http://dx.doi.org/10.1007/978-3-662-44199-2_88.

[Koh14c] Andrea Kohlhase. "Search Interfaces for Mathematicians". In: *Intelligent Computer Mathematics 2014*. Conferences on Intelligent Computer Mathematics. (Coimbra, Portugal, July 7–11, 2014). Ed. by Stephan Watt et al. LNCS 8543. Springer, 2014, pp. 153–168. URL: http://arxiv.org/abs/1405.3758.

[Koh14d] Michael Kohlhase. "A Data Model and Encoding for a Semantic, Multilingual Terminology of Mathematics". In: *Intelligent Computer Mathematics 2014*. Conferences on Intelligent Computer Mathematics. (Coimbra, Portugal, July 7–11, 2014). Ed. by Stephan Watt et al. LNCS 8543. Springer, 2014, pp. 169–183. URL: http://kwarc.info/kohlhase/papers/cicm14-smglom.pdf.

[Koh14e] Michael Kohlhase. "Mathematical Knowledge Management: Transcending the One-Brain-Barrier with Theory Graphs". In: *EMS Newsletter* (June 2014), pp. 22–27. URL: http://www.ems-ph.org/journals/newsletter/pdf/2014-06-92.pdf.

[KP13] Michael Kohlhase and Corneliu Prodescu. "MathWebSearch at NTCIR-10". In: *NTCIR Workshop 10 Meeting*. Ed. by Noriko Kando and Kazuaki Kishida. Tokyo, Japan: NII, Tokyo, 2013, pp. 675–679. URL: http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings10/pdf/NTCIR/MATH/04-NTCIR10-MATH-KohlhaseM.pdf.

[MH] *MathHub.info: Active Mathematics*. URL: http://mathhub.info (visited on 01/28/2014).

[Wat+14] Stephan Watt et al., eds. *Intelligent Computer Mathematics*. Conferences on Intelligent Computer Mathematics. (Coimbra, Portugal, July 7–11, 2014). LNCS 8543. Springer, 2014.