mArachna: Entwicklung von Wissensrepräsentationsmechanismen für die Mathematik

Sven Grottke (sfs@math.tu-berlin.de)
Sabina Jeschke (sabina@math.tu-berlin.de)
Nicole Natho (natho@math.tu-berlin.de)
Sebastian Rittau (rittau@math.tu-berlin.de)
Ruedi Seiler (seiler@math.tu-berlin.de)

Abstract: Die automatische Extraktion von Wissen aus natürlichsprachlichen Texten ist eine große technische Herausforderung, die – betrachtet man die Gesamtheit aller möglichen schriftlichen Quellen – heute noch als weitgehend ungelöst gelten muss. Wissenschaftliche und insbesondere mathematische Texte zeichnen sich jedoch durch einen höheren Grad der Strukturiertheit aus, und sie verfolgen stets das Ziel, Wissen zu transportieren und zu vermitteln. Mathematische Texte besitzen zudem in weiten Teilen eine starke Binnengliederung in "Bausteine" wie Definitionen, Theoreme etc., die als die Hauptträger des mathematischen Wissens aufgefasst werden können. Diese Textbausteine besitzen wiederum spezielle innere Textstrukturen, die einer computerlinguistischen Analyse zugänglich sind.
In diesem Artikel stellen wir ein System (mArachna) vor, das mathematische Zusammenhänge aus Texten extrahiert und in eine Wissensbasis integriert. Aus dieser Wissensbasis werden dann – als Antwort auf individuelle Abfragen – verschiedene Ausschnitte des mathematischen Wissens durch XML Topic Maps visualisiert. Ziel ist dabei insbesondere die Vermittlung von Übersichtswissen über das Wissensgebiet der Mathematik, sowie die Darstellung der innerfachlichen Zusammenhänge der mathematischen Objekte und Konzepte.

1 Hintergrund

Informationen und Wissen sind zentrale Begriffe in unserer heutigen Gesellschaft. Durch zahlreiche Publikationen, Bücher und nicht zuletzt durch das Internet entsteht eine Informationsflut, die durch einzelne Personen nicht mehr zu bewältigen ist. Darüber hinaus ist die manuelle Verarbeitung von Informationen sehr zeitaufwendig. Daher ist es sinnvoll, und zudem eine große technische Herausforderung, Mechanismen zu entwickeln, die Wissen aus natürlichsprachlichen Texten automatisiert extrahieren. Eine solche automatische Extraktion führt über die computerlinguistische Analyse von natürlichsprachlichen Texten bis zur Klassifikation und Visualisierung des enthaltenen Wissens.

Die künstliche Intelligenz und die Psychologie geben wichtige Impulse für die Klassifikation von Wissen [And01]. Diese Wissenschaften entwickeln Wissensklassifikationsmechanismen, wie z. B. semantische Netze, assoziative Netze und Knowledge Maps, die eine sinnvolle und effektive Organisation und Modellierung von Wissen ermöglichen und damit

auch für Anwendungen im Bereich des eLearning, eTeaching und eResearch von großer Bedeutung sind. Wissen wird hier als Relationen zwischen Aussagen und einzelnen Begriffen dargestellt. Durch diese Form der Darstellung von Wissen wird die Zugänglichkeit und die Verarbeitung von Wissen mittels des Computers wesentlich erleichtert.

2 Grundidee

mArachna ist ein System für die Extraktion von Wissen aus natürlichsprachlichen mathematischen Texten. Dabei verwendet mArachna bekannte Methoden der Computerlinguistik und der Wissensrepräsentation (z. B. Chomsky-Grammatiken). Ein Einsatzszenario des mArachna-Systems ist die Entwicklung eines user-adaptiven mathematischen Retrieval-Systems für die eLearning-Plattform Mumie [Mum, Jes04], die insbesondere für die mathematische Grundstudiumsausbildung von Ingenieuren an Hochschulen entwickelt wurde.

Der Content-Bereich der Mumie besteht im Kern aus – aktuell deutschsprachigen – mathematischen feingranularen Textbausteinen (und zusätzlich aus verschiedenen visualisierenden und interaktiven Elementen, die wir in dieser Arbeit im Folgenden jedoch vernachlässigen wollen). Wir bezeichnen diese speziellen Textbausteine als *Entitäten*. Entitäten drücken mathematische Objekte und Konzepte aus und entsprechen in Lehrbüchern einzelnen Definitionen, Theoremen, Beweisen usw. Diese Entitäten werden durch eine Ontologie verwaltet und durch sie in ein komplexes Netzwerk eingebunden, so dass die einzelnen Bausteine miteinander in Relation stehen. Die konzeptionelle Spezifikation der Mumie-Ontologie basiert dabei auf den Strukturen des mathematische Theoremgebäudes selbst, stellt also eine "intrinsische Fachontologie" dar.

mArachna stellt Mechanismen für die Erzeugung von Netzwerken für das Retrieval sowie für die Navigation auf diesen Netzwerken bereit. Als Quellen können dabei sowohl Wissensbausteine der Mumie-Plattform als auch Lehrbuchinhalte bestimmter Formate dienen. Diese Netzwerke spiegeln die kontextuellen Zusammenhänge wider. Sie sind eng an den natürlichsprachlichen mathematischen Text gebunden und können Zusammenhänge sehr feingranular abbilden, aber auch zur Erstellung von Übersichten verwendet werden. Damit ergeben sich verschiedene *Detaillierungsgrade* mit unterschiedlicher Informationstiefe.

Der Bezug auf die unterschiedlichen Bedürfnisse der Anwender für Retrieval- und Navigationsmechanismen ist ein wichtiges Konzept im mArachna-Projekt. Auf der Basis von nutzerdefinierten Anfragen, die zunächst über eine Anzahl von Auswahlmechanismen und Keyword-Suchen realisiert werden, werden Wissensnetze dynamisch on-the-fly generiert. Künftige Weiterentwicklungen der mArachna-engine sollen es ermöglichen, auch natürlichsprachliche mathematische Anfragen des Nutzers zu verarbeiten. Ebenso wird angestrebt, durch Einbeziehung des Nutzerprofils die Ausgabe an den Wissensstand des Nutzers anzupassen.

3 Linguistischer Aspekt

Die Entitäten sind Hauptinformationsträger in mathematischen Texten. mArachna verwendet ein linguistisches Klassifikationsschema [Jes04, Nat05], das die Basis liefert, um computerlinguistische Methoden auf die Entitäten anwenden zu können. Dieses Schema besteht aus den folgenden vier Ebenen (Abb. 1): Die Entitätenebene beschreibt die Zusammenhänge zwischen den verschiedenen Entitäten. Die Binnenstrukturebene charakterisiert die interne Textstruktur der Entitäten (z. B. Voraussetzungen und Aussage in Theoremen). Die Satzebene beschreibt die typischen Satzkonstruktionen in der mathematischen Sprache. Die unterste Ebene stellt die Wort- und Symbolebene dar, die die einzelnen Symbole und Wörter und auch die Zusammenhänge zwischen beiden schematisiert [NGJS05].

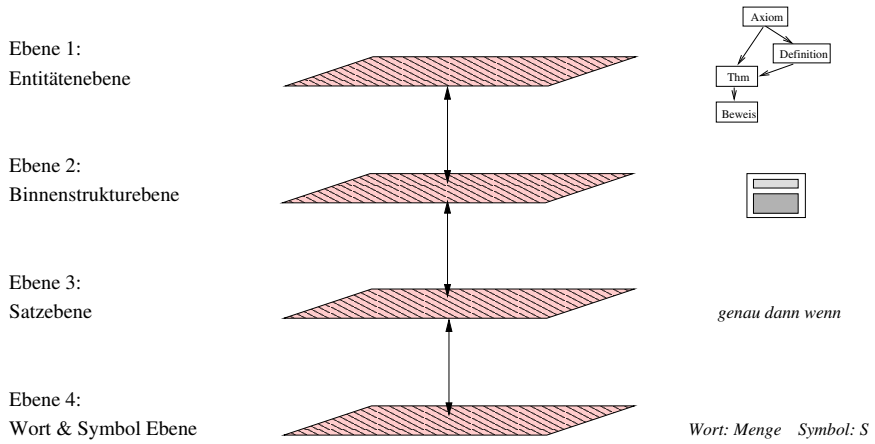

Abbildung 1: Sprachebenen der mathematischen Sprache

Unter Ausnutzung der durch dieses Schema gegebenen Strukturen und linguistischen Zusammenhänge werden mathematische Informationen aus den Texten extrahiert und in eine Wissensbasis integriert. Diese Wissensbasis besteht aus einer Graphenstruktur, die ihrerseits wiederum als eine Ontologie aufgefasst werden kann und in der Web Ontology Language OWL codiert ist [W3Cb], die ihrerseits auf dem RDF-Standard (Resource Description Framework) basiert [W3Ca].

Aus der computerlinguistischen Analyse der Entitäten (Informationen der Ebenen 2-4) erhalten wir Tripel [Hel00], die aus zwei Knoten, die die mathematischen Begriffe und Aussagen repräsentieren, und einer Relation, die die Knoten miteinander verbindet, bestehen; dabei repräsentieren verschiedene Typen von Relationen typische Sprachkonstruktionen oder spezielle Schlüsselwörter in den mathematischen Texten (Bsp.: Knoten entsprechend zwei Aussagen A und B, mit der Relation “ist äquivalent zu”). Durch die enge Anlehnung an die Sprachstruktur entsteht eine feingranulare Struktur der Wissensbasis.
Die Relationen zwischen den Entitäten (Ebene 1) werden mit Hilfe von Topic Map Technologie [Top] dargestellt. Diese Übersichten werden durch Extraktion von relevanten Informationen aus der Wissensbasis generiert, wobei insbesondere auch die Informationen

der zugrundegelegten Fachontologie (vergl. Kap. 2) verwenden werden.

4 Wissensbasis

Um die aus den mathematischen Texten extrahierten Informationen zu organisieren, werden sie in eine Wissensbasis eingebaut. Die Wissensbasis besitzt dabei zunächst ein elementares Grundwissen der Mathematik [Bou74], das sich aus der axiomatischen Mengenlehre und der Prädikatenlogik erster Stufe zusammensetzt.

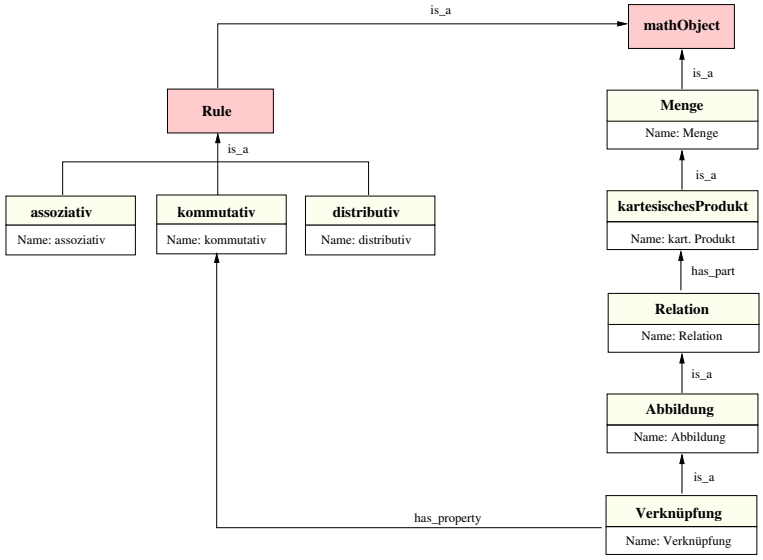

Abbildung 2: Elementare Strukturen in der Wissensbasis

Neu gewonnenes Wissen wird sukzessive in die existierende Wissensbasis integriert. Um Inkonsistenzen in der Wissensbasis zu vermeiden, wird ein semiautomatischer Ansatz verfolgt: Informationen aus mathematischen Texten werden nur dann in die Wissensbasis eingefügt, wenn die einzubauenden Informationen an existierendes Wissen – Knoten und Relationen – "angedockt" werden können und wenn keine widersprüchlichen Einträge oder Doppeleinträge entstehen. Im Falle fehlender Relationen und Knoten oder Inkonsistenz muss der Benutzer zusätzliche Informationen in die Wissensbasis einfügen oder die entsprechenden Konflikte beseitigen. Dieses "duale Modell" der Informationsorganisation ist nicht nur pragmatisches Hilfsmittel des mArachna-Systems, sondern entspricht auch den bekannten Modellen menschlicher Wissensverarbeitung: Menschen können Wissen nur dann sinnvoll in ihr Weltbild integrieren, wenn die gegebenen Informationen einen Bezug zu ihrem Vorwissen haben. Nicht ausreichendes oder fehlerhaftes Vorwissen kann zur Fehlinterpretation des neuen Wissens führen. Entsprechend wird falsches Wissen unter gewissen Bedingungen wieder entfernt bzw. korrigiert [And01].

5 Realisierung

Ein einfaches Beispiel veranschaulicht die Funktionsweise von mArachna. Dabei handelt es sich um eine Definition aus der linearen Algebra:

Definition 1 *Es gibt ein $e \in G$ mit den folgenden Eigenschaften:*

1. $a * e = a$ *für alle* $a \in G$
2. *Zu jedem $a \in G$ gibt es ein $a' \in G$ mit $a * a' = e$. $a' \in G$ heißt inverses Element von a.*

$e \in G$ heißt neutrales Element.

mArachna erkennt anhand des einleitenden Schlüsselwortes "Definition", dass es sich bei dieser Entität um eine Definition handelt, und wird dies im weiteren Lauf der Analyse berücksichtigen. Betrachten wir nun den letzten Satz genauer: *$e \in G$ heißt neutrales Element.*

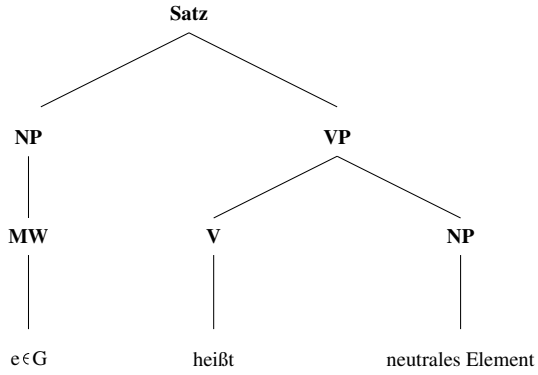

Abbildung 3: Schematische Darstellung der syntaktischen Analyse

Nach einer morphologischen Analyse, bei der einzelne Wörter erkannt werden, zerlegt die syntaktische Analyse den Satz in eine Nominalphrase (NP) und eine Verbalphrase (VP). Die Verbalphrase wird wiederum in ein Verb (V) und eine Nominalphrase unterteilt. Die Nominalphrase $e \in G$ besteht nur aus einer mathematischen Formel (MW), die während der syntaktischen Analyse nicht weiter zerlegt wird (Abb. 3).

Die nachfolgende semantische Analyse identifiziert "*heißt*" als Prädikat, "$e \in G$" als Subjekt und "*neutrales Element*" als Objekt. Da bekannt ist, dass es sich um eine Definition handelt, können das definierende und das definierte Element durch das Schlüsselwort "*heißt*" eindeutig identifiziert werden (Abb. 4). Zudem wird "$e \in G$" als typisches Konstrukt erkannt, um e als einen Bezeichner für ein Element von G einzuführen. Dies entspricht einer *is_element*-Relation zwischen e und G. Da e lediglich ein Bezeichner ist,

kann das definierte Objekt “neutrales Element” dafür substituiert werden. Das Ergebnis ist ein Tripel *(is_element, neutrales Element, G)*, das in die Wissensbasis eingefügt wird.

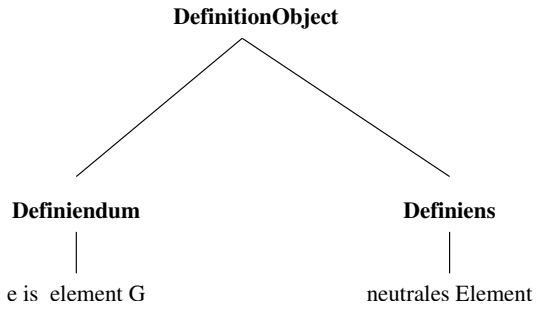

Abbildung 4: Schematische Darstellung der semantischen Analyse

Abb. 5 zeigt, wie das Tripel zur Wissensbasis hinzugefügt wird. Die Knoten “*Gruppe*” und “*inverses Element*” sind bereits vorhanden, und durch eine *is_element*-Relation verknüpft. Um das Tripel *(is_element, neutrales Element, G)* einzutragen, wird zunächst für “*neutrales Element*” und “*G*” überprüft, ob sie bereits als Knoten in der Wissensbasis existieren. Existieren beide nicht, dann wird ein Fehler ausgegeben. Sind beide Knoten vorhanden, und die verbindende Relation steht im Widerspruch zu der neu anzulegenden, dann wird ebenfalls ein Fehler ausgegeben. Im betrachteten Beispiel tritt keiner der beiden Fälle auf, und der neue Knoten “*neutrales Element*” wird angelegt und durch die *is_element*-Relation mit dem existierenden Knoten “*Gruppe*” verknüpft.

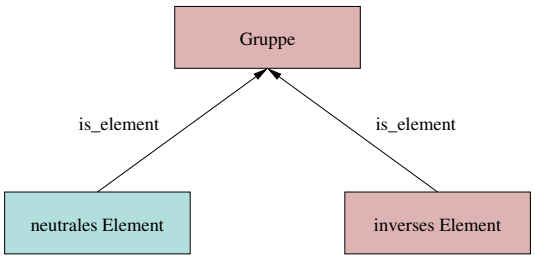

Abbildung 5: Ausschnitt aus der Wissensbasis

6 Wissensrepräsentionen

Auf der Grundlage der oben beschriebenen Wissensbasis konstruiert mArachna graphische Wissensrepräsentationen. Diese Repräsentationen liefern übersichtliche Darstellungen von Teilbereichen der Mathematik in verschiedenen Detaillierungsgraden. Die Visualisierung erfolgt als interaktiver Graph mit der Möglichkeit, durch Auswahl von Objekten

oder Bereichen und durch Zoom-Mechanismen durch die mathematische Wissenslandschaft zu navigieren.

Zur Darstellung der Wissensrepräsentation werden XML Topic Maps [Top] verwendet: Topic Maps basieren auf der Metasprache XML und dienen der Organisation und Strukturierung von Informationen. Sie orientieren sich an den Wissenstrukturierungsmechanismen des Menschen[1], weshalb sich ihr Einsatz für das Information Retrieval System von mArachna anbietet. Für die XML Topic Maps existieren Anwendungsprogramme, wie z. B. das Java-basierte TM4J-Framework [tm4], das auch die Technologie der Mumie-Plattform unterstützt.

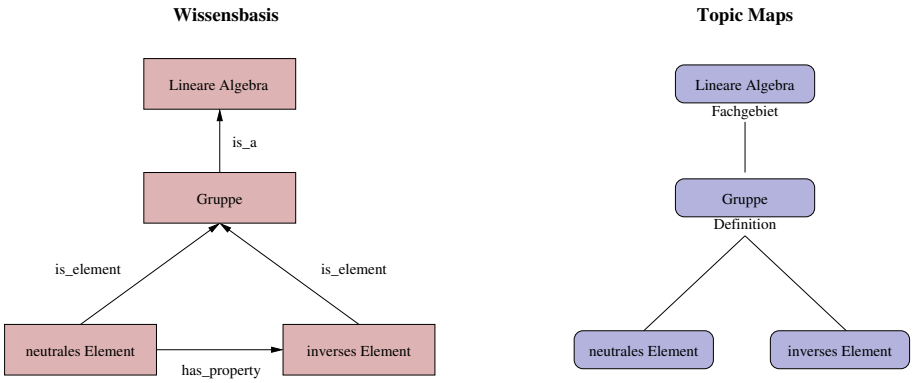

Abbildung 6: Relationen in der Wissensbasis (links) – Darstellung in den Topic Maps (rechts)

Durch eindeutige Definitionen von Begriffen und Sachverhalten sowie die wohlstrukturierten Beschreibungsmechanismen mathematischer Sachverhalte eignen sich XML Topic Maps insbesondere sehr gut zur Darstellung von mathematischem Wissen. Dabei kann die Struktur der Wissensbasis und damit der verarbeiteten mathematischen Texte gut auf die Topic Maps-Struktur übertragen werden.

XML Topic Maps extrahieren die für ihre Erzeugung relevanten Informationen aus der Wissensbasis und stellen deshalb geeignete, dem menschlichen Verständnisprozess angepasste, Auswahlen des dort verwalteten Wissens dar. Sie erlauben die effizientere und übersichtlichere Darstellung von Informationen und ihren Zusammenhängen – und sie ermöglichen darüber hinaus benutzerspezifische Sichtweisen der betrachteten Wissensausschnitte, indem Rahmenbedingungen (z. B. scopes) an das zu visualisierende Gebiet formuliert werden.

[1] Dagegen orientiert sich RDF mehr an der Metadaten-Strukturierung aus Sicht möglichst effizienter Verarbeitung durch den Computer [Gar]. mArachna verwendet beide Standards: die Wissensbasis basiert auf RDF/OWL, das Information Retrieval operiert auf einem Spiegelbild der Wissensbasis und basiert auf Topic Maps.

7 Überblick und Ausblick auf das mArachna-Projekt

Das mArachna-Projekt befindet sich derzeit in einer Phase prototypischer Implementation. Der Prototyp demonstriert – auf ausgewählten Textbausteinen – die Durchführbarkeit des in diesem Text beschriebenen halbautomatischen Ansatzes, um semantische Informationen aus typischen mathematischen Textelementen – Entitäten – zu extrahieren. Diese Informationen können gespeichert und in die Wissensbasis integriert werden. Darauf basierend können einige Topic Maps dargestellt werden.

Als nächster Schritt ist die Integration von mArachna in die Mumie Plattform [Mum] geplant, um den Prototypen im praktischen Einsatz zu testen. Von Interesse sind dabei das Verhalten bei großen Datenmengen und deren Handhabbarkeit in Bezug auf die feingranulare Struktur der Wissensbasis einerseits, Fragestellungen der Benutzerführung und Useradaption andererseits. In linguistischer Hinsicht wird als nächster Schritt die Analyse der englischen mathematischen Sprache angestrebt.

Literatur

[And01] J. R. Anderson. *Kognitive Psychologie*. Spektrum Akademischer Verlag, Heidelberg, Berlin, 3rd. Auflage, 2001.

[Bou74] N. Bourbaki. Die Architektur der Mathematik. In M. Otte, Hrsg., *Mathematiker über die Mathematik*. Springer, Berlin, Heidelberg, New York, 1974.

[Gar] L.M. Garshol. Living with topic maps and RDF. `http://www.ontopia.net/topicmaps/materials/tmrdf.html`.

[Hel00] H. Helbig. *Die semantische Struktur natürlicher Sprache. Wissensrepräsentation mit MultiNet*. Springer, Berlin, Heidelberg, 2000.

[Jes04] S. Jeschke. *Mathematik in Virtuellen Wissensräumen – IuK-Strukturen und IT-Technologien in Lehre und Forschung*. Dissertation, Technische Universität Berlin, Berlin, April 2004.

[Mum] Mumie community. Mumie. `http://www.mumie.net`.

[Nat05] N. Natho. *mArachna: Eine semantische Analyse der mathematischen Sprache für ein computergestütztes Information Retrieval*. Dissertation, Technische Universität Berlin, Berlin, Februar 2005.

[NGJS05] N. Natho, S. Grottke, S. Jeschke und R. Seiler. mArachna: A Classification Scheme for Semantic Retrieval in eLearning Environments in Mathematics. Proceedings of the 3rd International Conference on Multimedia and ICTs in Education, June 7-10, 2005, Careres/Spain, June 2005.

[tm4] tm4j.org. TM4J – Topic Maps 4 Java. `http://tm4j.org`.

[Top] TopicMaps.org. Topic Maps. `http:/www.topicmaps.org`.

[W3Ca] W3C. Resource Description Framework (RDF). `http://www.w3.org/RDF/`.

[W3Cb] W3C. Web Ontology Language (OWL). `http://www.w3.org/2004/OWL/`.