

Managing mathematical texts with OWL and their graphical representation

Sabina Jeschke, Marc Wilke
University of Stuttgart
RUS - Center of Information Technologies
sabina.jeschke@rus.uni-stuttgart.de
marc.wilke@iits.uni-stuttgart.de

Nicole Natho, Olivier Pfeiffer
Berlin University of Technology
MuLF - Center for Multimedia
in Education and Research,
{natho, pfeiffer}@math.tu-berlin.de

Abstract

Mathematical knowledge contained in scientific digital publications poses a challenge for intelligent retrieval mechanisms. Many current approaches use statistical (e.g. Google) or natural language processing methods to find correlations in texts and annotate texts semantically. However both kinds of approaches face the problem of extracting and processing knowledge from mathematical equations.

The presented system is based on natural language processing techniques, and benefits from characteristic linguistic structures defined by the language used in mathematical texts. It accumulates extracted information snippets from texts, symbols, and equations in knowledge bases. These knowledge bases provide the foundation for the information retrieval.

This article describes the concepts and the prototypical technical implementation.

1. Introduction

Numerous books and publications are released day-to-day imparting knowledge. Particularly the World Wide Web is a giant unsystematic knowledge conglomeration. Additionally too much information induces the so-called “information glut” that is not easy to handle. Therefore knowledge management takes a significant position in modern organizations and society. To face up to the new challenges, applications to specialize and generalize information gain in importance.

This article describes the knowledge management system KEA that extracts mathematical information from different kind of sources such as textbooks and publications, and integrates the information into knowledge bases in a suitable format. Many approaches of information extraction and accumulation are based on statistical methods, e.g. the search algorithm implemented by Google, looking for correlations between notions or concepts.

However, these approaches do not take the precision of the mathematical language into account.

Natural language processing (NLP) methods like “Head-driven Phrase Structure Grammar” has the possibility for accurate automatic semantic annotation of texts to extract mathematical “fine-grained” correlations between terms and concepts. Such a system provides a lot of new challenges: automated knowledge acquisition and automated generation of ontologies of mathematical fields, new requirements of user interfaces, new search and retrieval mechanisms, and different visualization models of the extracted mathematical knowledge, etc. As a result, mathematical knowledge management systems necessitate sophisticated mechanisms for accumulation, storage, merging, evaluation, and representation of mathematical information for different kinds of applications like encyclopedias, context sensitive library search systems, intelligent book indexes and e-learning software.

KEA implements the above listed features of mathematical knowledge management systems. It integrates extracted information represented as elements of mathematical concepts and objects as well as relations and connections between them into knowledge bases. This generation is possible because of the characteristic structure of the mathematical language, characterized by representative text elements (“entities”) such as definitions, theorems and proofs that are generally used to describe mathematical objects and concepts. Entities are the keys to access knowledge within mathematical texts. Propositional logic is reflected in the syntax and semantic structure of mathematical texts facilitating the extraction of information. In addition, each author has his own style to write mathematical texts. The information contained in mathematical texts forms complex relationship models in the shape of networks that define ontologies of the analyzed texts.

The basic ontology that forms the skeletal mathematical structure for all ontologies are based on the ideas of Bourbaki [1] and Hilbert [2]: mathematics as a whole can be derived from a small set of axioms using propositional logic. (This approach is valid within this context, despite Gödel's Incompleteness Theorem [3], since existing and already proven mathematical knowledge is mapped into a knowledge base whereas adding new (unproven) theorems and checking the (in-)consistency of theorems is not intended.)

Therefore, new mathematical knowledge is integrated on the basis of the skeleton ontology, the inherent existing natural language structure in mathematical texts itself and the natural language structure implemented by the author. The extracted mathematical objects and concepts are annotated with additional information regarding their role in the original texts.

The resulting knowledge bases can be used to create an overview of mathematical content. To manage these knowledge bases, the Jena framework [4] is used. To retrieve information we use the search engine Apache Lucene [5].

2. Related Works

To date we have not heard of a project similar to KEA regarded as a turn-key solution. However, KEA consists of different building blocks, which can be compared with related projects. Basic results on mathematical ontologies were accomplished by Gruber and Olsen [6]. MBASE [7] is an example of a manually written mathematical ontology. Investigating English mathematical texts is accomplished by Baur [8]. An example of a manually written mathematical encyclopedia is MathWord [9].

WordNet [10, 11] and GermaNet [12] are intelligent thesauri providing a number of suitable synonyms and definitions for given queried expressions similar to the semantic analysis of KEA. Concepts for automated semantic analysis of the German language were performed by Helbig [13].

Mizar [14, 15, 16] is used for describing mathematical proofs through a formal human and machine-readable language by an automated reasoning system.

DIALOG [17] processes natural language for mathematical validation in an automated reasoning system. HELM [18] uses metadata provided by authors for semantic annotation of mathematical texts used in digital libraries. In

combination with HELM MoWGLi [19] is a retrieval interface based on pattern-matching within mathematical equations and logical expressions.

3. Technical Design

KEA consists of three separate modules (see figure 1):

- preliminary analysis,
- NLP analysis, and
- knowledge base.

The technical design associates to the following general specifications:

- **Modularization:** Each component is an independent module that can easily be replaced.
- **Standard internal data formats:** All data is represented in standard data formats (TEI, MathML, XML, RDF/OWL).
- **L^AT_EX as input format:** L^AT_EX was chosen as input format since it is the standard format for publications in mathematics and the natural sciences.
- **Easy-accessible control of the NLP analysis:** All rules governing the syntactic and semantic analysis should be modifiable by non-programmers to facilitate extensions.

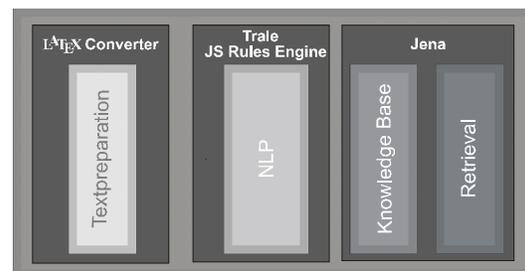


Fig. 1. Modules of KEA

3.1 Preliminary and NLP Analysis

Entities are the main carriers of information in mathematical texts. Based on a linguistic classification scheme [20, 21, 22], entities are analyzed using NLP techniques. The scheme has four levels (see figure 2):

- **Entity Level:** entities and their arrangement within the text
- **Structure Level:** the internal structure of an entity (e.g. the proposition and assumptions of a theorem)
- **Sentence Level:** characteristic sentence

structures, frequently found in mathematical texts

- **Word and Symbol Level:** single symbols and words and the relations between them [21].

Using these structures and linguistic relations mathematical information is educed from texts and integrated into knowledge bases through a complex NLP analysis. The resulting knowledge bases consist of directed graphs, representing objects and concepts, and their interrelations. They are encoded in the standardized web ontology language OWL [23] and based upon the above described skeleton ontology.

The NLP analysis of entities produces triples of two nodes (representing mathematical objects and relations) and one relation (describing types of connection between the nodes). The triples are integrated into the knowledge bases. The triple structure is complex and very fine-grained (i.e. nodes are nodes of other triples), and different kinds of relations describe different kinds of keywords or phrases within analyzed texts (e.g. two nodes representing two theorems, connected with the relation “is equivalent to”).

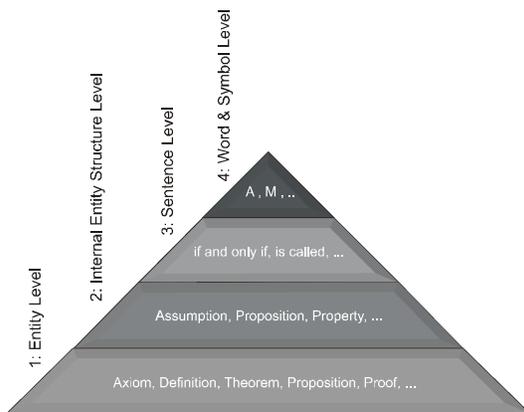


Fig. 2. Linguistic Classification Scheme

For the internal representation of the analyzed texts, the KEA system uses TEI (Text Encoding Initiative [24]). Mathematical symbols and equations are encoded in Presentation MathML [25]. For this purpose the KEA provides a L^AT_EX-to-TEI/MATHML-converter.

The preliminary analysis is implemented in Java using rule-based string comparisons: Entities are segmented into single sentences annotated with some meta-information (their parent entity and their role within the internal structure of that entity), and equations are

separated from the enveloping text (replaced by placeholders).

The NLP analysis manipulates every extracted single natural language sentence by using the TRALE-System (a PROLOG adaptation [26] of ALE for German [27]) based on a Head-Driven Phrase Structure Grammar. The underlying dictionary and grammar of TRALE have been expanded to include the particularities of the mathematical language, in order to provide a comprehensive syntactic and some semantic information. The result is an abstract syntax tree, reflecting the structure of the analyzed sentences. Now, all equations, that have been separately syntactically processed, have to be reintegrated for further semantic analysis. At this state, we process simple built-on equations. Techniques for analyzing complex equations are developed, and are partly implemented.

The semantic analysis is implemented as an embedded JavaScript interpreter [28]. Syntax trees are categorized according to characteristic structures of the entities. Each category is transformed into the specific triple structure defined by external JavaScript rules. These rules reflect typical mathematical language constructs of the corresponding basic mathematical concepts (e.g. proposition, assumption, definition of a term etc.). The resulting triples are annotated with additional information from the original text.

The resulting triple structures are OWL-documents satisfying the OWL-DL standard, and partly the OWL-FULL standard. The problem is, that for each triple element it has to be decided if it represents OWL classes or individuals of OWL classes.

3.2 Equations Analysis

Symbols, abbreviations and equations are the most important elements in mathematical texts. They present a key source to gather information. It is necessary to be able to analyze their content during the NLP analysis. Currently, only simple symbols and equations can be processed. However, we are developing an approach to improve our skills for analyzing mathematical equations and symbols. Therefore, the use of a syntactical analysis for equations will be made available, similar to those used in computer algebra systems in combination with contextual grammars (e.g. Montague grammars) to correlate the information given in an equation with information already provided in the surrounding

text. We would like to point out that our main goal is not a machine-based understanding of mathematical information. Therefore we try to avoid the use of automatic reasoning systems. Instead, equations and symbols are treated as a different representation of mathematical knowledge to be integrated into the knowledge base similar to the use of the natural language texts.

3.3 Knowledge Management

In the KEA System it is necessary that each analyzed text creates its own knowledge base, and therefore its own ontology, by reason of each author having his own preferences, spelling styles and idiosyncrasies of the representation of a mathematical field. For these knowledge bases, an (“upper”) ontology describing general concepts of all texts consists of all information of general mathematical language constructs (“is called”, “Let ...”, etc.) exists. Because of the speech diversity of different authors also (“domain”) ontologies exist within the knowledge bases. Domain ontologies describe the different usage of styles of phrasing and notation, or the author’s didactical preferences. A new and empty knowledge base consists of fundamental mathematical knowledge including axiomatic field theory and first order logic (cf. figure 3).

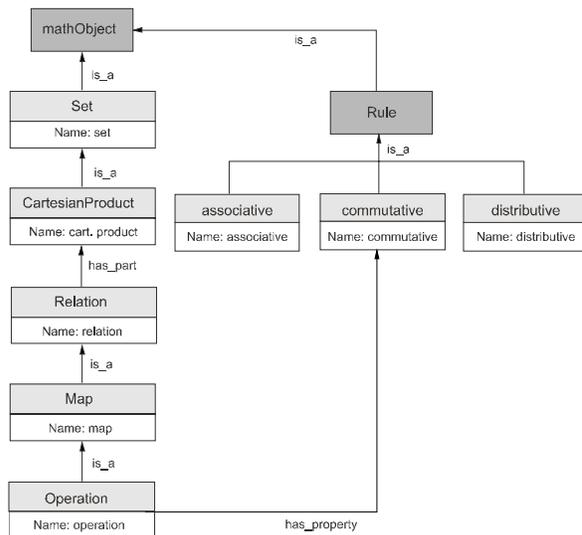


Fig. 3. Basic Structure of the Knowledge Base

To be obvious, the sheer amount of data is a big challenge for retrieving information from the

knowledge bases. To encounter this problem, the knowledge base is developed within the Java-based Jena Framework [28]. This open source framework consists of several useful APIs for generation and manipulation of the knowledge base with RDF (Resource Description Framework), OWL (Web Ontology Language [22]), and RDFS (RDF Scheme). Generally, RDF forms triples to create statements. Several implementations of Jena interfaces represent RDF basic elements (resources, literals, statements, etc.). Additionally, Jena can read and write different representation forms (M-Triples or N3), and different mechanisms of RDF-models (primary memory or/and miscellaneous data bases). Particularly Jena provides an RDF query language SPARQL [29] which is oriented on SQL as interference engine to derive additional RDF assertions.

To avoid inconsistencies within the knowledge base, new information is added using as required a semi-automated application. Information can only be integrated into the knowledge base if and only if there are no conflicts or contradiction with existing entries. Consequently, this means that new nodes have to be connected to existing ones whereas duplications are also not allowed. A semi-automatic approach means that the system administrator manually eliminates such conflicts by adding additional information or deleting the conflicting entries using a web-based interface. The basic idea of this approach is the well-known model of human knowledge processing in psychology: integration of new knowledge by humans is only possible if it can be linked to existing knowledge conformed by the persons world view. Consequently, correct knowledge has to be newly interpreted. A short evaluation shows that conflicts do not occur very frequently, and thus they are not the rule. The semi-automated administrator intervention is performed through a comfortable web interface manipulating the Jena Java classes.

However, the ideal way of extracting information from the knowledge base does not seem to be obvious currently. Not only the sheer amount of data is a problem, but also the complex structure of the combined triples and the additional attached information induce non-easily accessible knowledge bases. Therefore it makes sense to regard the different language preferences of the authors, and eliminate them or use them early. One idea is to store all textual exchangeable abbreviations and symbols separately in text files. Then we categorize

different linguistic levels like the axiomatic structure or internal structure (i.e.: the exclusive representation of definitions). From this level we can zoom into deeper nesting. It is also possible to use the natural structure of the text itself like chapters, subchapters, paragraphs, etc. In addition, we can track user's searching results and techniques within our knowledge bases to imitate their structural searching routines. Because of the problem of the sheer amount of data and complex information processing, the categorization of knowledge bases is our main focus of development.

3.4 Merging

The merging process is one part of our efforts to categorize our knowledge bases. On the one hand all the knowledge bases of the different authors to one mathematical field shall be unified. In this process some difficulties are hidden as even in mathematics, different authors use different types of definitions and theorems. The differences can be so subtle that even a human reader needs profound considerations. On the other hand, tools and strategies for unitizing the unified knowledge bases into smaller knowledge bases have to be developed and realized. Such smaller knowledge bases can be accommodated to the preferences and requirements of specialized user groups. Moreover the unitized knowledge bases can be used as the basis of an intelligent retrieval system based on the concept of personal information agents (PIA). For example, we are using different versions of entities for elementary or high schools. Moreover unitized knowledge bases using different notations could contribute to the harmonization of teaching material within courses.

3.5 Web Retrieval Interface

Currently a web interface for information retrieval from KEA is being implemented. The interface pursues the basic philosophy of the "Personalized Home" of iGoogle, and follows up the metaphor of a desktop within a browser, allowing a high adaptability to the needs of the individual user. The web-based desktop is implemented using JavaScript and XML-technologies (Ajax applications), offering predefined widgets that can easily be adapted to the preference of the user, provided in form of a substantial, free library. These widgets for

example include: advanced search tools, personalised histories of queries and results, including social tagging, graphical representation mechanisms of different views (low-level and upper ontologies, different entity levels, etc.) on the knowledge base generated via Graphviz [30]. We use Ajax technologies to show how it is possible to combine semantic web technologies and new generations of web-based applications like wikis and folksonomies, in a suggestive and mutually supportive way.

4. Evaluation and Outlook

KEA is capable of analyzing German texts. Currently KEA is expanded to analyze English texts as well. The current implementation proves the operability of the semi-automated method of semantic extraction for special text elements. The ontology mapping approach shall be evaluated in a next step by analyzing whole textbooks on Linear Algebra and Analysis with a subsequent merging of the resulting knowledge bases and a concluding comparison of the generated outcome with some mathematical "standard" encyclopedia – respectively the corresponding part of the used encyclopedia.

For further evaluation purposes the system shall be connected to an appropriate cooperative knowledge space platform for mathematics and natural sciences.

References

- [1] Bourbaki, N.: Die Architektur der Mathematik. Mathematiker über die Mathematik. Springer, Berlin, Heidelberg, New York (1974)
- [2] Hilbert, D.: Die Grundlagen der Mathematik. Abhandlungen aus dem mathematischen Seminar der Hamburgischen Universität 6 (1928) 65–85
- [3] Gödel, K.: Über formal unentscheidbare Sätze der Principia Mathematica und verwandte Systeme I. Monatsheft f. Mathematik und Physik (1931-1932) 147f
- [28] Jena: A Semantic Web Framework for Java. <http://jena.sourceforge.net> (last visited: 09/15/2007)
- [5] The Apache Software Foundation: Apache Lucene. <http://lucene.apache.org/> (last visited: 09/15/2007)
- [6] Gruber, T., Olsen, G.: An Ontology for Engineering Mathematics. Technical Report KSL-94-18, Stanford University (1994)

- [7] M. Kohlhase, A. Franke, "MBase: Representing Knowledge and Context for the Intergration of Mathematical Software Systems.", *Journal of Symbolic Computation* 23:4, pp. 365 - 402 (2001)
- [8] Baur, J.: Syntax und Semantik mathematischer Texte. Master's thesis, Universität des Saarlandes, Fachbereich Computerlinguistik (November 1999)
- [9] Wolfram Research: MathWorld. <http://mathworld.wolfram.com> (last visited: 09/15/2007)
- [10] Fellbaum, C.: WordNet: An Electronic Lexical Database. MIT Press, Cambridge, London (1998)
- [11] Fellbaum, C.: WordNet. <http://wordnet.princeton.edu> (last visited: 09/15/2007)
- [12] GermaNet Team: GermaNet. <http://www.sfs.uni-tuebingen.de/lsd/english.html> (last visited: 09/15/2007)
- [13] Helbig, H.: Knowledge Representation and the Semantics of Natural Language. Springer, Berlin, Heidelberg, New York (2006)
- [14] Urban, J.: MoMM – Fast Interreduction and Retrieval in Large Libraries of Formalized Mathematics. *International Journal on Artificial Intelligence Tools* (15(1)) (2006) 109–130
- [15] Urban, J.: MizarMode – An Integrated Proof Assistance Tool for the Mizar Way of Formalizing Mathematics. *Journal of Applied Logic* (2005)
- [16] Urban, J.: XML-izing Mizar: Making Semantic Processing and Presentation of MML Easy, MKM2005 (2002)
- [17] Pinkall, M., Siekmann, J., Benz Müller, C., and Kruijff-Korbyova, I.: DIALOG. <http://www.ags.uni-sb.de/~dialog/> (last visited: 09/15/2007)
- [18] Asperti, A., Padovani, L., Sacerdoti Coen, C. and Schena, I.: HELM and the Semantic Web. In Boulton, R.J., Jackson, P.B., eds.: *Theorem Proving in Higher Order Logics*, 14th International Conference, TPHOLS 2001, Edinburgh, Scotland, UK, September 3-6, 2001, Proceedings. Volume 2152 of *Lecture Notes in Computer Science*, Springer (2001)
- [19] Asperti, A., Zacchioli, S.: Searching Mathematics on the Web: State of the Art and Future Developments. In: *Joint Proceedings of the ECM4 Satellite Conference on Electronic Publishing at KTH Stockholm, AMS - SM M Special Session, Houston / (2004)*
- [20] Jeschke, S.: *Mathematik in Virtuellen Wissensräumen - IuK-Strukturen und IT-Technologien in Lehre und Forschung*. PhD thesis, Technische Universität Berlin (2004)
- [21] Natho, N.: MARACHNA: Eine semantische Analyse der mathematischen Sprache für ein computergestütztes Information Retrieval. PhD thesis, Technische Universität Berlin (2005)
- [22] Grottko, S., Jeschke, S., Natho, N., Seiler, R.: 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, Caceres/Spain (2005)
- [23] W3C: Web Ontology Language. <http://www.w3c.org/2004/OWL> (last visited: 09/15/2007)
- [24] The TEI Consortium: Text Encoding Initiative. <http://www.tei-c.org> (last visited: 09/15/2007)
- [25] W3C: MathML. <http://www.w3.org/Math> (last visited: 09/15/2007)
- [26] Müller, S.: TRALE. <http://www.cl.uni-bremen.de/Software/Trale/index.html> (last visited: 09/15/2007)
- [27] Müller, S.: *Deutsche Syntax deklarativ: Head-Driven Phrase Structure Grammar für das Deutsche*. In: *Linguistische Arbeiten*, No. 394. Max Niemeyer Verlag, Tübingen (2005)
- [28] Mozilla Foundation: Rhino. <http://www.mozilla.org/rhino/> (last visited: 09/15/2007)
- [29] W3C: SPARQL. <http://www.w3.org/TR/rdf-sparql-query/> (last visited: 09/15/2007)
- [30] AT&T Research: Graphviz. www.graphviz.org/ (last visited: 09/15/2007)