Machine-learning methods for classification and content authority in mathematics software

UDC Seminar
Lisbon
2015-10-29

Ulf Schöneberg (FIZ Karlsruhe)
Wolfram Sperber (FIZ Karlsruhe)

FIZ Karlsruhe
Leibniz-Institut für Informationsinfrastruktur

Agenda

- Background and motivation
- MSC and controlled vocabulary
- Key phrase extraction
- Classification
- About the mathematical language
- SMGloM – a special authority tool for mathematics
- Summary

The background and motivation

- idea of reviewing journals ("Jahrbuch über die Fortschritte der Mathematik“, 1868):
  give the mathematicians an (complete) overview about the progress in mathematics
- former role of mathematical reviewing journals:
  "memory of the mathematical community"
- increasing number of mathematical publications
  (1868: 876 items ; 2010: 107,204 items)
  → reviewing journals are under permanent development:
  - new methods for content analysis were used:
    key phrases, classification scheme
    classification schemes used in mathematics:
    Mathematics Subject Classification (MSC2010)
    Math Reviews, zbMATH
    UDC (Referativni Journal " Matematika ")

MSC (I)

Home | MSC Wiki | Feedback | MSC2010 TiddlyWiki | CD Contents

Mathematics Subject Classification MSC2010

The Mathematics Subject Classification (MSC) has undergone a general revision, with some additions and changes, and corrections of existing errors, thus creating MSC2010 as a successor to the previous MSC2000. Mathematical Reviews (MR) and Zentralblatt für Mathematik (Zbl) have carefully considered all feedback and used it in preparing their joint MSC revision.

The Final MSC2010 revision was made public here in May 2009, and deployed in production in July 2009 by MR for MathSciNet and Zbl for ZMATH. A final Third Public Working Draft was finished in July 2008, a Second Public Working Draft in March 2008, and the First Public Working Draft was published in January 2008. There was an interactive TiddlyWiki form of the Third Public Working Draft of the MSC2010.

The drafts were publicly developed using a MediaWiki at this site, namely the MSCwiki. This will remain open for public view. PDF forms of the current MSC and derived lists are there. There is also available an interactive TiddlyWiki version of the MSC2010. On this site one may also view the contents of a CD of the Final Public Working Draft distributed at the North American Joint Winter Mathematics Meetings in Washington DC, 5-9 January 2009.

Comments and suggestions may still be made using an interactive form, but only simple errors can now be fixed. Larger issues will have to be considered later as a possible revision to a succeeding MSC2020 takes place.

www.msc2010.org

MSC(II)

- hierarchical scheme:
  (63, 528, 5606 classes on the top, the second, and the third level)
- strong overlapping (different kinds of similarity → semantic relations between classes)
- only a rough definition of the content of the MSC classes by
  - class labels
    and
  - the position within the classification scheme
- periodic updating
- formalization (SKOS scheme:
  http://msc2010.org/resources/MSC/2010/MSC2010)

The (un)controlled vocabulary of zbMATH

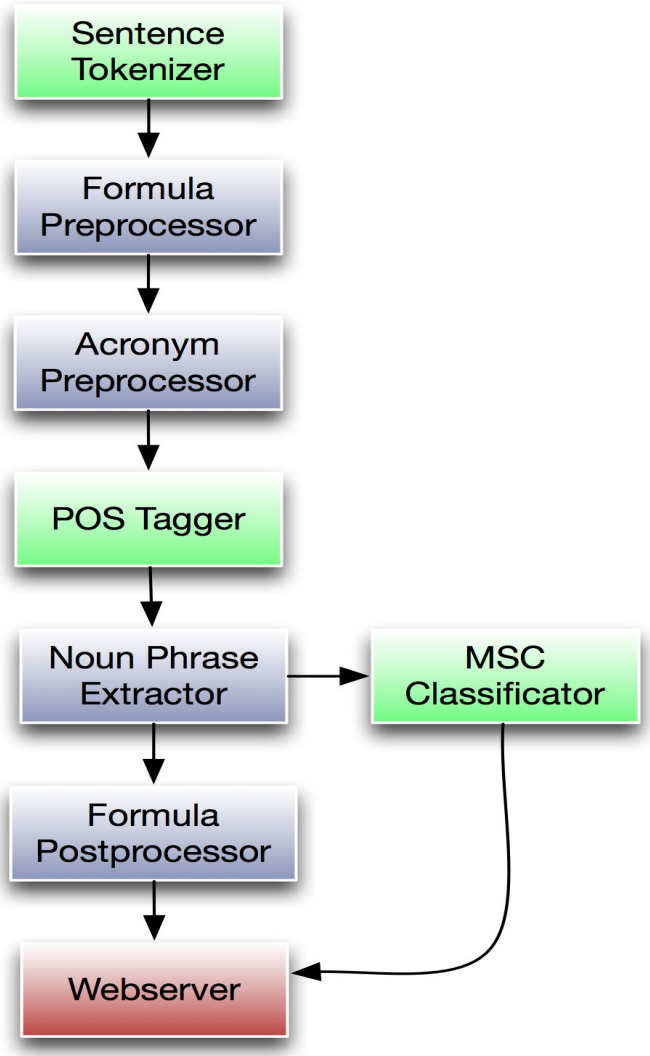
- authors often use keywords for a short characterization of the content
- zbMATH provides keywords since the 60s
- keywords in zbMATH are (un)controlled terms! (created by authors, reviewers, editors)
- Observations
  - keywords are not keywords but really key phrases
  - zbMATH: ~ 3,500,000 items, ~ 9,100,000 classfication codes, ~ 10,000,000 (not disjunct) key phrases
  - 'semi-standardization' of key phrases: often the names of the MSC classes are used as keywords,
  - often key phrases contain not more information than the MSC code

Idea:

- key phrase extraction by NLP methods
- automatic classification by using key phrases

Special problem: Symbols and formulae

Workflow for key phrase extraction and classification



FIZ Karlsruhe
Leibniz-Institut für Informationsinfrastruktur

Key phrase extraction (I)

Methods
NLP for extracting key phrases in zbMATH data

- First step: Tokenization (tokens are separated by blanks, deleting of special characters, e.g., dots, hyphens)

- Second step: Preprocessing
  Preprocessing of formulae (symbols and formulae are encoded in TeX in zbMATH,
  hence, symbols and formulae can be identfified
  and processed in a separate way)
  Preprocessing of acronyms (acronyms are identified and are substituted by full form)

Key phrase extraction (II)

- Second step: POS Tagging
  POS tagging: marking the syntactic role of each token (word)
  Penn Treebank POS scheme is used: 45 tags
  Stanford POS Tagger
  symbols and formulae are typed as nouns (NN)

Use of Stanford's dictionary of the common English language
Bulding up specialized dictionaries:

- resolution of acronyms
- proper names (extension of Stanford's dictionary:
  name of mathematicians or special mathematical terms)

Key phrase extraction (III)

- third step: Noun phrase extraction
  Noun phrases are typical for key phrases
  searching for noun phrases
  Definition of characteristic patterns for noun phrases,
  e.g.,
  Knaster-Kuratowski-Mazurkiewicz lemma $\K3L$
  NNP NNP NNP NN NNP

Key phrase extraction (IV)

Up to now we have extracted noun phrases (this set contains noun phrases

Fourth step: relevant noun phrases
different methods:

- scoring of noun phrases (manually and automatically)
- neural networks
- comparing phrases with existing mathematical encyclopediae Wikipedia, Encylopedia of Mathematics, PlanetMATH, SMGloM , ...

Key phrase extractor

Lipschitz stability of solutions of linear-quadratic parabolic control problems with respect to perturbationsThis paper deals with the optimal control of linear parabolic equations for quadratic cost functionals. The author studies the stability of optimal solutions to control problems with respect to perturbations of the state equation and the cost functional. He first proves a stability result for the L^2-perturbations when a coercivity assumption is satisfied. Next, using a bootstrap argument he establishes L^∞-stability of solutions for L^∞-perturbations The analysis mainly relies on regularity results for parabolic equations. This kind of result plays a major role in the convergence analysis of the SQP method [see, for example, F. Tröltzsch, SIAM J. Control Optimization 38, No. 1, 294-312 (1999; following review)].

---

unknown words:

bootstrap	noun
coercivity	noun
functionals	common noun, plural
Tröltzsch	proper noun, singular
perturbationsThis	proper noun, singular
294-312	numeral cardinal
linear-quadratic	adjective
Optimization	proper noun, singular
Mathematics	proper noun, singular
perturbations	common noun, plural

sequential quadratic programming method	1
linear-quadratic parabolic control problems	1
Applied Mathematics J. Control	1
results for parabolic equations	1
L^∞-stability of solutions	1
equations for quadratic cost	1
stability of optimal solutions	1
kind of result	1
problems with respect	1
Society for Industrial	1
stability of solutions	1
state equation	1
F. Tröltzsch	1
author studies	1
perturbationsThis paper	1
coercivity assumption	1
stability result	1
convergence analysis	1
major role	1
optimal control	1
bootstrap argument	1

zurück

Use of key phrases for classification (I)

Further step: Classification
methods of automatic text classification used:

- Naive Bayes classificators,
- Support Vector Machines (SVM),
- C4.5 trees,
- and combinations of these methods

basing on

- key phrases

alternatively

- zbMATH 'full texts' (abstracts)

Use for classification (II)

- The classification quality (precision, recall) basing on noun phrases is higher than with full texts..
- But, the quality is strongly depending from the subject (MSC classes).
  Automatic classification works fine for classes which have a minor overlapping with other classes
  Automatic classification makes problems for classes
  with major overlapping.
  (remark: also the vocabulary is overlapping for these classes)

Key word extraction by neural networks

Classical machine-learning methods in text processing:
Bag-of-words model (tokens and its frequencies)

(Convolutional recurrent) neural networks use not only single words but analyze also the context → semantic approach
training set is the base for learning, its quality is essential
example: semantic similar words in the English Wikipedia (631 Mio tokens)
NN method provides amazing results

Open source tool for neural networks: word2vec (Google)

Use of neural networks in zbMATH

blue	positive	linear	prime number	algebra	color
red	nopnnegative	nonlinear	primes	ring	pixel
green	nonzero	quadratic	integers	module	texture
colored	$k>0$	bilinear	square-free	K-algebra	image
monocromatic	bounded	parametric	cardinality	C^*-algebra	luminance
2-coloring	$\alpha>0$	differential	number theory	subalgebra	RGB

Use of neural networks in zbMATH (II)

Remarks:
- Input are tokens or phrases
- Some similarities seem to be 'non-trivial'.

Neural networks methods in text processing - when it works?
- terminology must be homogeneous (no metaphors, no" lyrics")
- zbMATH data are nearly perfect data for neural networks
  the subjects are (relatively) clear, no metaphors are used
- we need good training data

→ one strategy: building up a high-quality training set for mathematics and using neural networks
but what is with formulae?

Some remarks about the mathematical language?

- Mathematics is a natural language but with some specialities
- Mathematical language is dual: Mathematical concepts, objects, and models can be represented by terms and symbols (notations).
- Names (of terms) / notations are ambigous: Different names / notations can be used for the same mathematical concept, object or model.
- Names / notations can be used for various mathematical concepts, objects or models.
- Names / names can have different linguistic / notational forms. Normalization (canonical forms) for authority control .
- Terms and their notations are given by one or more definitions. (The equivalence of definitions must be proved.)

SMGloM – a terminological and notational base for mathematics

Therefore, we have developed a new concept for a semantic knowledge base (and authority tool) for the mathematical language: SMGloM

SMGloM: acronym for Semantic Multilingual Glossary of Mathematics
https://mathhub.info/mh/glossary

shortly: SMGloM contains mathematical terms (canonical forms) given by a definition, their (semantified) notations of a mathematical concept, object or model plus the relations to other mathematical terms.

Navigation

Help

Report issue

User login

Username *

Password *

- Create new account
- Request new password

Log in

smglom

One of the challenging aspects of mathematical language is its special terminology of technical terms that are defined in various mathematical documents. The SMGloM is a lexical resource that combines the characteristics of dictionaries and glossaries with those of mathematical ontologies. It facilitates a large variety of knowledge management applications without requiring full formalization, the cost of which would be prohibitive. See the license here.

Responsible: m.kohlhase@jacobs-university.de

Statistics

- Staging Ground Various mathematical concepts to be sorted into SMGloM repositories
- Sets Basic properties of sets
- Mathematical Vernacular The special language to express mathematical knowledge
- Elementary Calculus Terminology for the mathematical study of change.
- Mathematical Constants Special mathematical constants
- Number Theory (general) General terminology of Number Theory
- Mathematical Identities Equivalent terms of one or more variables
- Topology Terminology for connectedness, continuity, boundary, and the like.
- Geometry Terminology for shape, size, relative position of figures, and properties of space.
- Trigonometry Terminology about functions describing the relationships between lengths and angles in triangles.
- Linear Algebra Terminology for vector spaces, lines, planes, subspaces, matrixes, ...
- Functional Analysis Vector spaces with some kind of limit structure.
- Analysis Calculus, its applications and enhancements
- Special Numbers The objects of Number Theory.
- Prime Numbers Special numbers that are primes.
- Magic Squares Terminology about Magic Squares
- Number Fields Terminology about numbers, their representations, and their operations
- Elementary Algebra Elementary algebra encompasses some of the basic concepts of algebra, one of the main branches of mathematics.
- Numbertheoretical Functions Important functions of Number Theorie
- Elementary Graph Theory Graphs are structures that can model many things

Navigation

Help

Report issue

User login

Username *

Password *

- Create new account
- Request new password

Log in

SMGloM Glossary

ro en tr de zht zhs

- B -powersmooth Definition, Concept Graph de
- B -smooth Definition, Concept Graph de
- b base encoding Definition, Notations, Concept Graph de
- b base encoding Definition, Notations, Concept Graph de
- j th digit Definition, Notations, Concept Graph de
- k -ful number Definition, Concept Graph de
- k -full number Definition, Concept Graph de
- k -perfect Definition, Concept Graph de
- k -powerful number Definition, Concept Graph de
- k -rough number Definition, Concept Graph de
- k -simplex Definition, Concept Graph de
- n -fold composition Definition, Notations, Concept Graph zhs de
- n friendly -tuple Definition, Concept Graph de
- n th derivative Definition, Concept Graph de
- p -closure Definition, Concept Graph zhs de
- s -additive Definition, Concept Graph de
- a -equal Definition, Concept Graph de
- a -Modul Definition, Concept Graph
- a -module Definition, Concept Graph
- a -Ulam number Definition, Notations, Concept Graph de
- a -Ulam sequence Definition, Concept Graph de

Navigation

Help

Report issue

User login

Username *

Password *

- Create new account
- Request new password

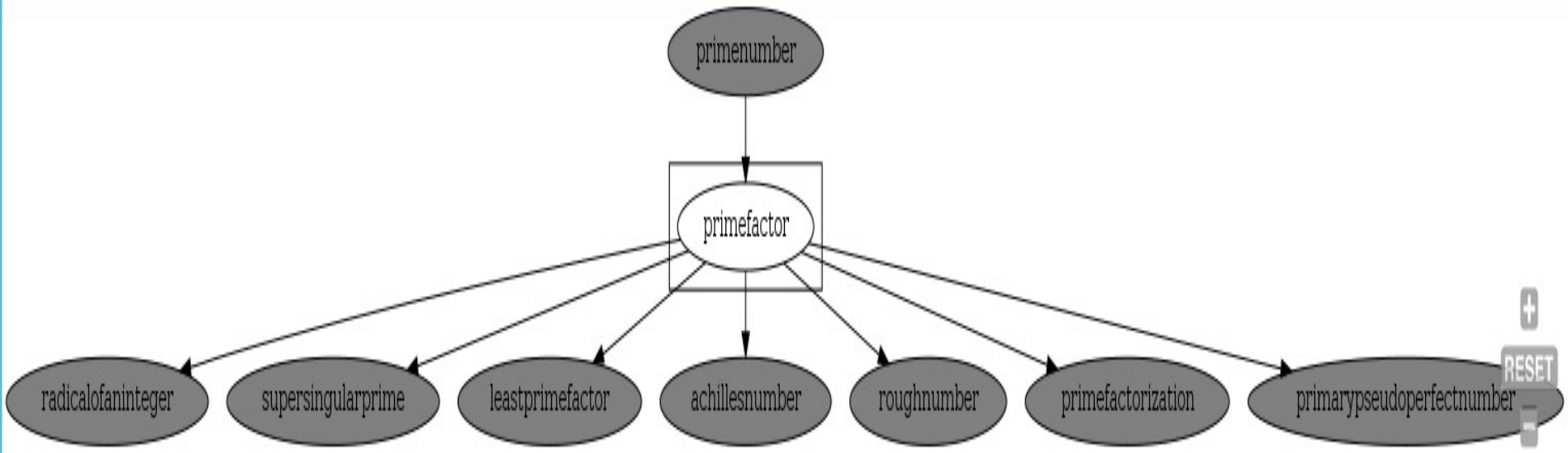
Log in

primenumber.en

View OMDoc Source SVG

- primenumber.en
  includes 2
  A **prime number** is a natural number greater than 1 that has no positive divisors other than 1 and itself.
  The number of prime numbers not greater than n is written as .

Conversion Succeeded (with warnings) Show Log



Semantic relations are presented as graphs

Summary

- Standardized methods of linguistics and computers science can also be used for text analysis in mathematics..
- But the mathematical language also requires the development of own concepts and methods reflecting the specifics of the mathematical language.
- New authority tools, e.g., a semantic glossary of mathematics are needed.

Thanks for your attention!