The DeLiVerMATH project
Text analysis in mathematics

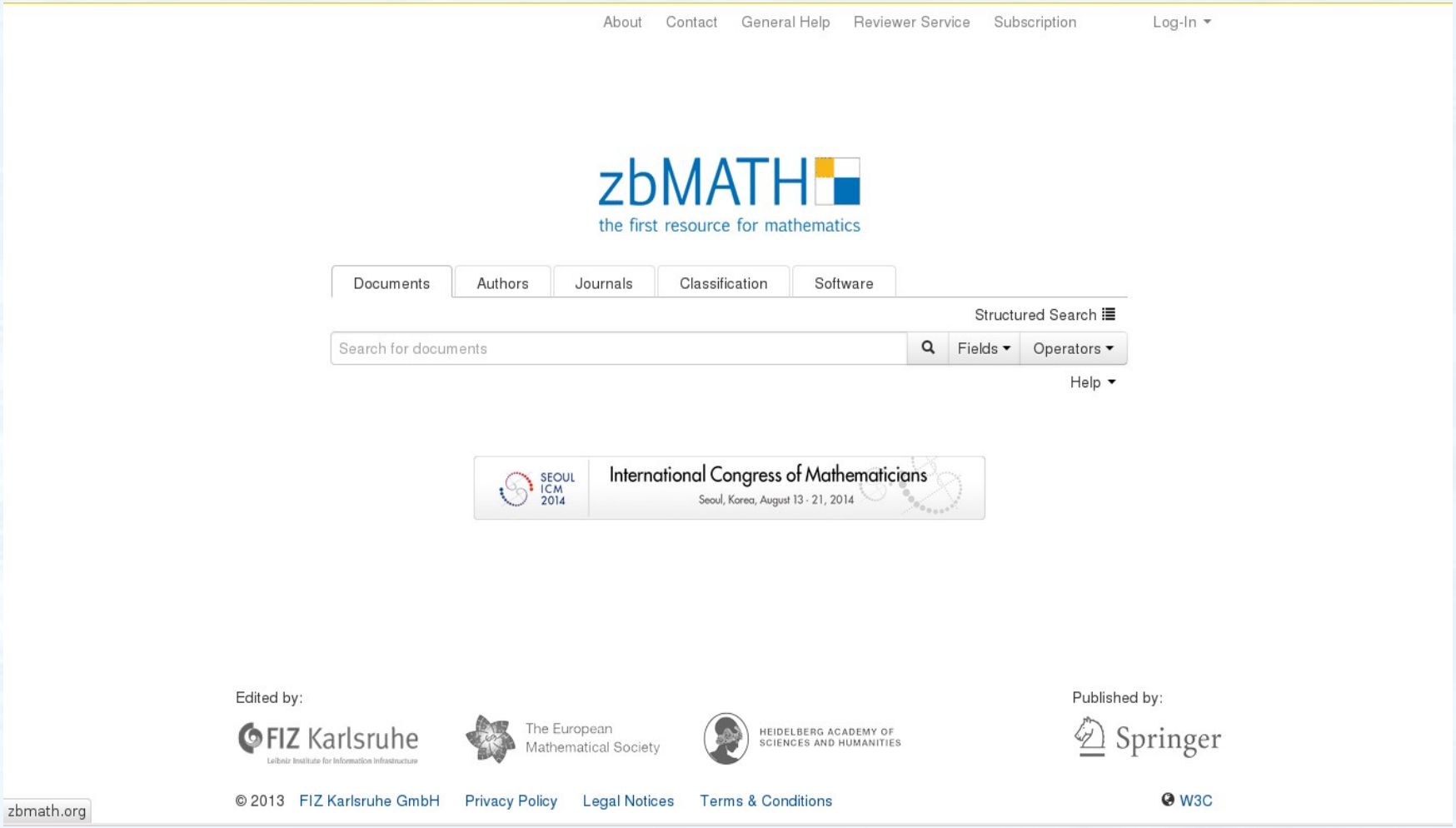
Ulf Schöneberg
Wolfram Sperber

CICM 2013
Systems & Projects
2013-07-09
Bath

Agenda

- Problem and state of the art analysis
- Concept
- Prototype
- Evaluation

Our database



Some facts

- worldwide production in mathematics and applications more than 100,000 peer-reviewed publications p.a., e.g., 2008: 105,324 items in zbMATH (2013-07-02)
- size of mathematics relevant literature is increasing, especially in application areas
- new features:
  - author disambiguation,
  - author profiles,
  - references,
  - multilinguality
- content analysis (reviewing, key phrase extraction, classification) is mainly done manually

New machine-based concepts and methods for content analysis - basics

Key phrase extraction and a mathematical glossary

- Use of Natural Language Processing (NLP) techniques
  - Morphological analysis
  - PoS Tagging (word-category disambiguation
    different tagging schemes;
    Penn Treebank tag scheme)
  - Formal Grammars / syntactical analysis
- Dictionaries ('Brown corpus', more than 1,000,000 English words)

Penn Treebank Tagset

Here are the most important tags. See also:
M. Marcus, Beatrice Santorini and M.A. Marcinkiewicz: Building a large annotated corpus of English: The Penn Treebank. In *Computational Linguistics*, volume 19, number 2, pp313-330.

POS Tag	Description	Example
CC	coordinating conjunction	and
CD	cardinal number	1, third
DT	determiner	the
EX	existential there	*there* is
FW	foreign word	d'hoevre
IN	preposition/subordinating conjunction	in, of, like
JJ	adjective	green
JJR	adjective, comparative	greener
JJS	adjective, superlative	greenest
LS	list marker	1)
MD	modal	could, will
NN	noun, singular or mass	table
NNS	noun plural	tables
NNP	proper noun, singular	John
NNPS	proper noun, plural	Vikings
PDT	predeterminer	*both* the boys
POS	possessive ending	friend*s*
PRP	personal pronoun	I, he, it
PRP$	possessive pronoun	my, his
RB	adverb	however, usually, naturally, here, good
RBR	adverb, comparative	better
RBS	adverb, superlative	best
RP	particle	give *up*
TO	to	*to* go, *to* him
UH	interjection	uhhuhhuhh
VB	verb, base form	take
VBD	verb, past tense	took
VBG	verb, gerund/present participle	taking
VBN	verb, past participle	taken
VBP	verb, sing. present, non-3d	take
VBZ	verb, 3rd person sing. present	takes
WDT	wh-determiner	which
WP	wh-pronoun	who, what
WP$	possessive wh-pronoun	whose
WRB	wh-abverb	where, when

Adaptions to zbMATH data

- tokens outside of the Brown corpus ('semicontinuity')
- special notations ('fixed-point')
- names of mathematicians ('Lyapunov')
- acronyms ('LMI')
- formulae (TeX-encoded)
- preferenced sentence constructs ('…of...')

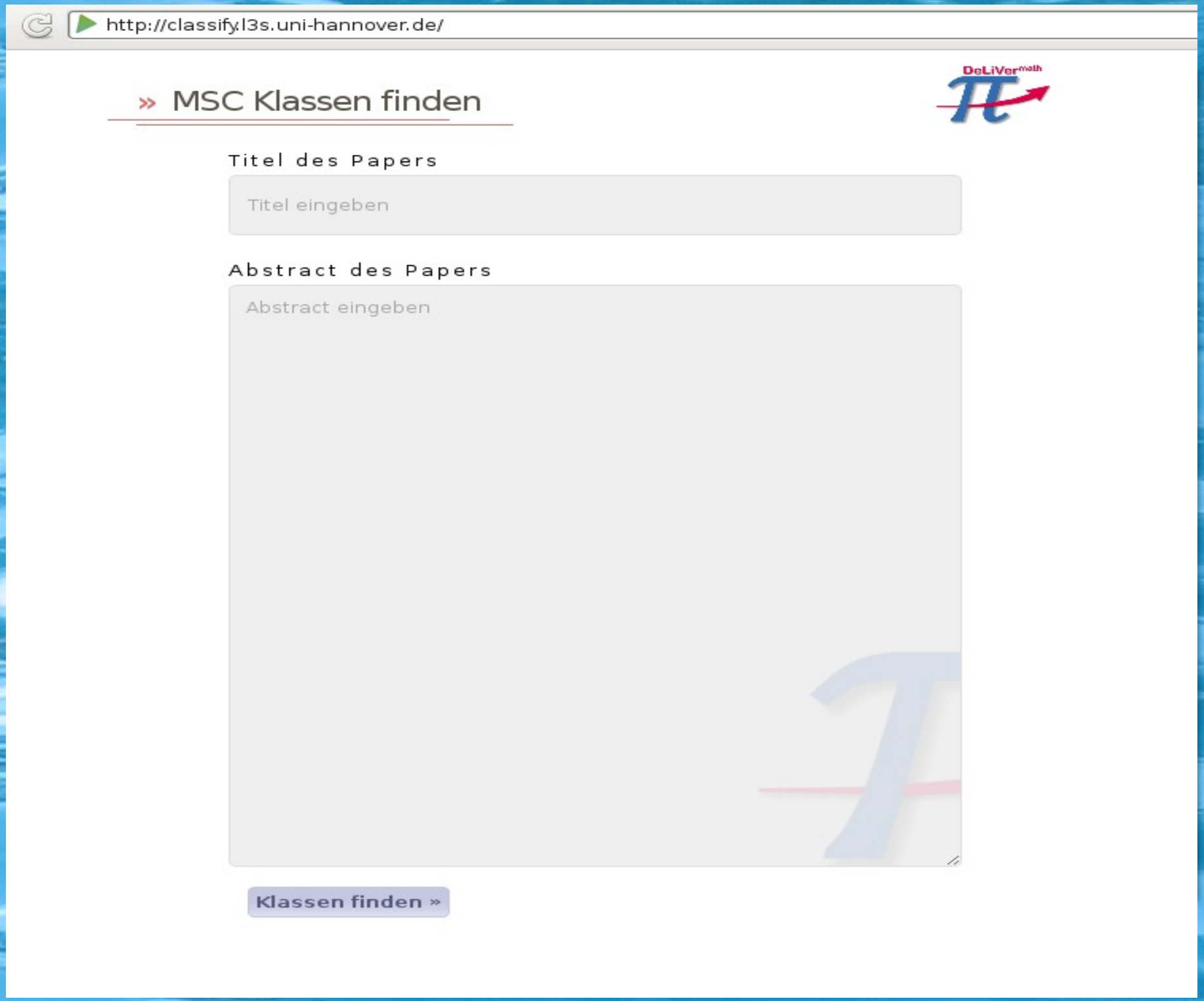
→ development of special dictionaries,
special treatment of formulae

Classification

- Implementation of standard tools :
  - Naive Bayes,
  - SVM
  - Random Forrest

It was done in parallel by L3S and Zentralblatt MATH.

L3S has used the complete abstracts for classification, Zentralblatt MATH extracted key phrases


http://classify.l3s.uni-hannover.de/
» MSC Klassen finden
DeLiVermath
Titel des Papers
Titel eingeben
Abstract des Papers
Abstract eingeben
Klassen finden »

Hybrid extended Fourier series for optimal control of nonlinear algebraic dynamical systems. The paper introduces a new method for finding optimal control of algebraic dynamic systems. The structure of algebraic dynamical systems is nonlinear with quadratic and bilinear terms. A new hybrid extended Fourier series is introduced, and state and control variables of the system are expanded by this series. Moreover, properties of new series are presented, and integration and product operational matrices are obtained. Using operational matrices, optimal control of the systems is converted to a set of simultaneous nonlinear algebraic relations. An illustrative example is included to compare our results with those in literatures.

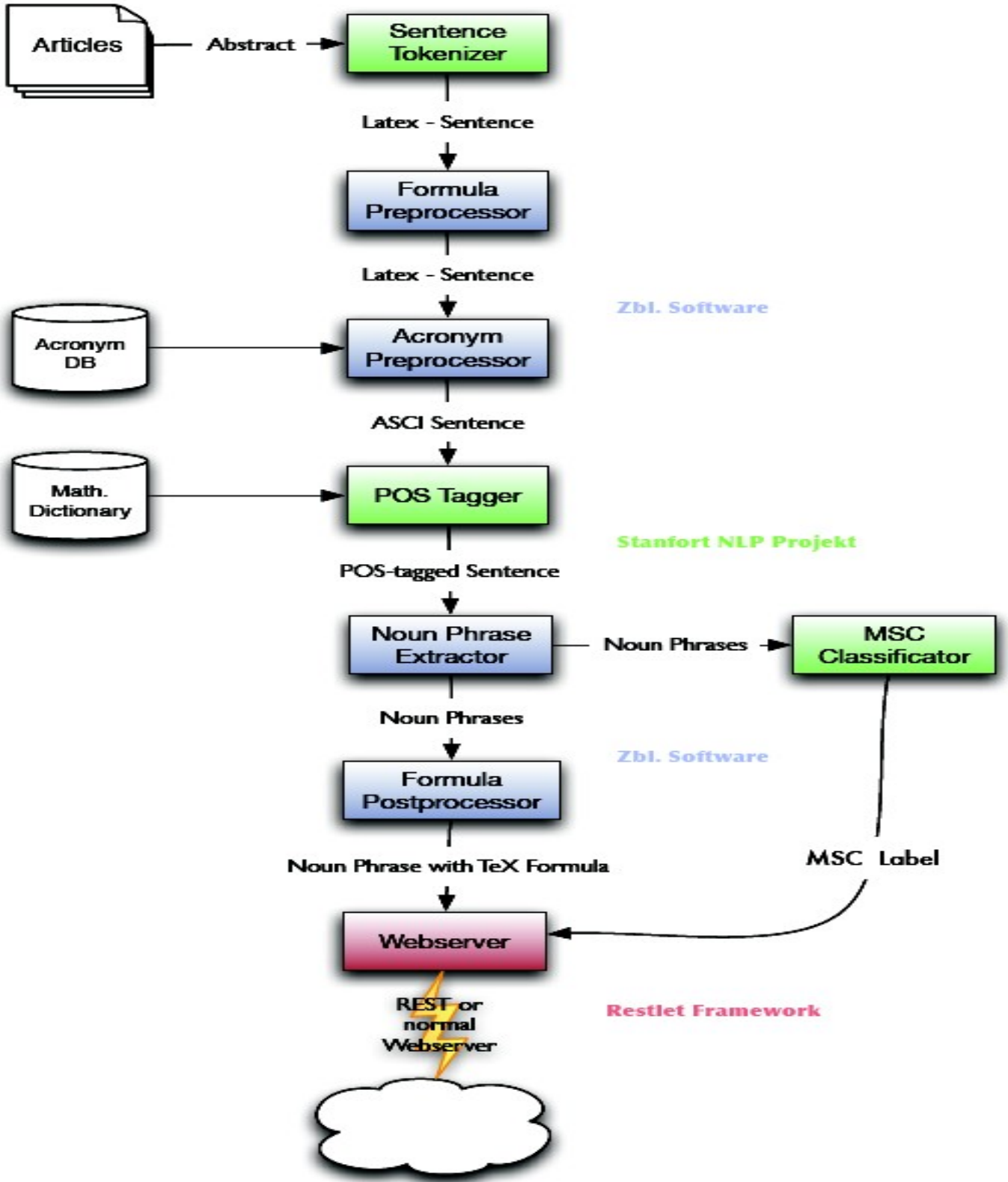
---

msc (sv): **93 65**

msc (matrix): 93C10

Unknown Words:
Hybrid noun

simultaneous nonlinear algebraic relations	1
nonlinear algebraic dynamical systems	1
nonlinear algebraic dynamic systems	
properties of new series	1
algebraic dynamical systems	1
product operational matrices	1
algebraic dynamic systems	1
quadratic systems	
optimal control	3
optimal control	
Fourier series	2
hybrid extended Fourier series	
bilinear terms	1
bilinear systems	
control variables	1
new hybrid	1
new method	1
illustrative example	1
operational matrices	1


Articles
Abstract
Sentence Tokenizer
Latex - Sentence
Formula Preprocessor
Latex - Sentence
Zbl. Software
Acronym DB
Acronym Preprocessor
ASCI Sentence
Math. Dictionary
POS Tagger
Stanfort NLP Projekt
POS-tagged Sentence
Noun Phrase Extractor
Noun Phrases
MSC Classificator
Noun Phrases
Zbl. Software
Formula Postprocessor
Noun Phrase with TeX Formula
MSC Label
Webserver
REST or normal Webserver
Restlet Framework

Features of the tools

- L3S tool
  only classification, but a fine granular classification (all three levels of the MSC)
- Zentralblatt MATH tool
  classification (but only on the top level)
  and key phrase extraction

  including a comparison with given key phrases and MSC classification

Evaluation - first results (I)

- classification
  standard measures: precision and recall or the F1 score
  comparison of manually and automatically generated MSC classifications marks (for each class)
  test set: ~ 300 items from the volume 1260 of zbMATH in the classes 03, 05C and 93

	Precision	Recall	F1-score
MSC 03	0.9255	0.8286	0.8744
MSC 05C	0.9103	0.9706	0.9395
MSC 93	0.9687	0.7209	0.8267

Evaluation – first results (II)

Some comments:

- The corresponding values of our tool are similar.
- The parameters on the second MSC level differ between different classes (the number of publications in the test set is to small).
- Up to now, the quality of manual classification was not evaluated (we could make a first evaluation by a cluster analysis and compare the MSC classes).

Evaluation – first results (III)

Key phrase extraction

- How we can measure the key phrases?
  - number of relevant key phrases (manual evaluation)
- Results: The number of key phrases is increasing. But, we have also a large number of non-significant phrases:
  - irrelevant phrases ('important theorems')
  - redundant phrases (more or less the same content)

Evaluation – first remarks (IV)

Some comments:

- Up to now, the tool produces to much key phrases (> 50 %). → manual work
- We need special tools to decrease the number of non-relevant key phrases, e.g., a special dictionary of irrelevant phrases, dictionaries with preference phrases and some kind of normalization, ...
- The improved tool will be used to create a Math glossary.

A first facit

- We are on the way but there a lot of things are to do.
- The tools show that machine-based procedures could support the content analysis of mathematical publications.
- Standard methods of NLP and classification can be used but must be enriched by special concepts which include the specific of mathematical language.