Classification of Multilingual Mathematical Papers in DML-CZ

Preliminary Excursion

Petr Sojka, Radim Řehůřek

Masaryk University, Faculty of Informatics, Brno, Czech Republic
sojka@fi.muni.cz, xrehurek@fi.muni.cz

Abstract. The growth of digital repositories of scientific documents is speed-ed up by various digitisation activities. Almost all papers of mathematical journals are reviewed by either Mathematical Reviews or ZentralBlatt Math, summing up to more than 2.000.000 entries.
In the paper we discuss possibilities and experiments we did on the data of Czech Digital Mathematics Library, DML-CZ with the goal of developing novel scalable methods of document classification and retrieval of multilingual mathematical papers.

1 Motivation – Project of Digital Mathematics Library

You always admire what you really don't understand. (Blaise Pascal)

Mathematicians from all over the world dream of World Digital Mathematics Library [1], where (almost) all of reviewed mathematical papers in all languages will be stored, indexed and searchable with the today's leading edge information retrieval machinery. A good resources towards this goals–in addition to the publisher's digital libraries–are twofold:

1. 'local' repositories of digitised papers as NUMDAM [2][1], DML-CZ [3][2] or born-digital archives CEDRAM [4][3]), arXiv.org>math[4]
2. two review services for the mathematical community: both ZentrallBlatt Math[5] and Mathematical Reviews[6] have more than 2.000.000 entries (paper metadata and reviews) from more than 2300 mathematical serials and journals.

Google Scholar[7] is becoming useful in the meantime, but lacks specialised math search and metadata guessed from parsing crawled papers are of low quality (compared to the controlled repositories).

Both review services agreed on the supported Mathematics Subject Classification (MSC) scheme[8], and currently used MSC 2000 is being revised for use in

[1] http://www.numdam.org [2] http://www.dml.cz [3] http://www.cedram.org
[4] http://arxiv.org/archive/math [5] http://www.zblmath.fiz-karlsruhe.de/MATH/
[6] http://www.ams.org/mr-database [7] http://scholar.google.com [8] http://www.ams.org/msc/

Petr Sojka, Aleš Horák (Eds.): Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2007, pp. 89–96, 2007. © Masaryk University, Brno 2007

2010 (MSC2010). Most journals request classification being used already by authors when submitting journals for publication; however, most of retrodigitised papers published before MSC 1990 are not classified by MSC in the databases.

Within the DML-CZ project we have investigated possibilities to classify (retrodigitised) mathematical papers by machine learning techniques, to enrich math searching capabilities and to allow semantically related search. As text of scanned pages is usually optically recognised, machine learning algorithms may use not only metadata (and reviews, if any), but also full text. Interesting question to pose is to find to which extent mathematical formulae are important for classification, document similarity measures, and search.

2 Data Preprocessing

> We run carelessly to the precipice, after we have put something before us to prevent us seeing it. (Blaise Pascal)

There are many modelling techniques for given classification task in the area of pattern recognition. To design a classifier, we have to choose measurable features. These features should be as discriminative as possible with regard to the pattern of interest. Most of the methods use bag of words representation of a document. There are methods such as Latent Semantic Analysis (LSA), that try to find main document topics based on word co-occurences in documents.

2.1 Primary data

The data available for experiments are metadata and full texts of mathematical journals covered by DML-CZ project. During the first three years of the project, we have digitized and collected data in digital library, accessible via web tool called Metadata editor[9]. To date (November 2007), in the digitised part there are 351 volumes of 9 journals: 1449 issues, 11725 articles on 173779 pages. We are promised to get another 15000+ full texts of articles of other digitisation project. In addition, digital born data of currently processed articles by various journals are being imported into the library, as workflow of paper publishing process was modified a bit so that all fine-grained metadata including the full text are exported for the digital library for long-term storage (CEDRAM).

By 2009, we target for a digital library with about 50000 mathematical articles with full texts, and much more with basic article metadata (abstracts, reviews).

For first experiments, we have used two types of data:

1. texts from scanned pages of digitized journals (usually before 1990, where no electronic data are available);
2. texts from 'digital-born' papers, written in TEX.

[9] `editor.dml.cz`

We started our experiments with retrodigitised articles from the *Czech Mathematical Journal* (CMJ)[10] from years 1951 to 1991 (starting 1992 there exist born-digital data). We took only those papers where both primary MSC classification in Zentrallblatt and Mathematical Reviews agree. This was done to ensure a clean training and evaluation set. In addition, we have used only part of the text corpus of the journal: only MSC categories with more than 60 papers were trained in the experiments. We got 925 papers in eight MSC categories:

class 05-xx (Combinatorics): 129 articles
class 06-xx (Order, lattices, ordered algebraic structures): 178 articles
class 08-xx (General algebraic systems): 64 articles
class 20-xx (Group theory and generalizations): 147 articles
class 34-xx (Ordinary differential equations): 146 articles
class 46-xx (Functional analysis): 70 articles
class 53-xx (Differential geometry): 87 articles
class 54-xx (General topology): 104 articles

Second text corpora we used in our experiments was created from papers of Journal *Archivum Mathematicum*[11] from years 1992–2007, where we had TEX source files available. For machine learning we use MSC categories for which we had at least 40 papers – they were categories 34, 53 and 58 (Global analysis, analysis on manifolds).

2.2 Preprocessing and methods used

It is widely known that design of the learning architecture is very important, as is preprocessing, learning methods and their parameters [5].

First part of the preprocessing is tokenizing the input documents. We used alphabetic tokenizer, with lowercase or Krovetz stemmer [6]. No stoplists were used, no word bi-grams, no lemmatization yet.

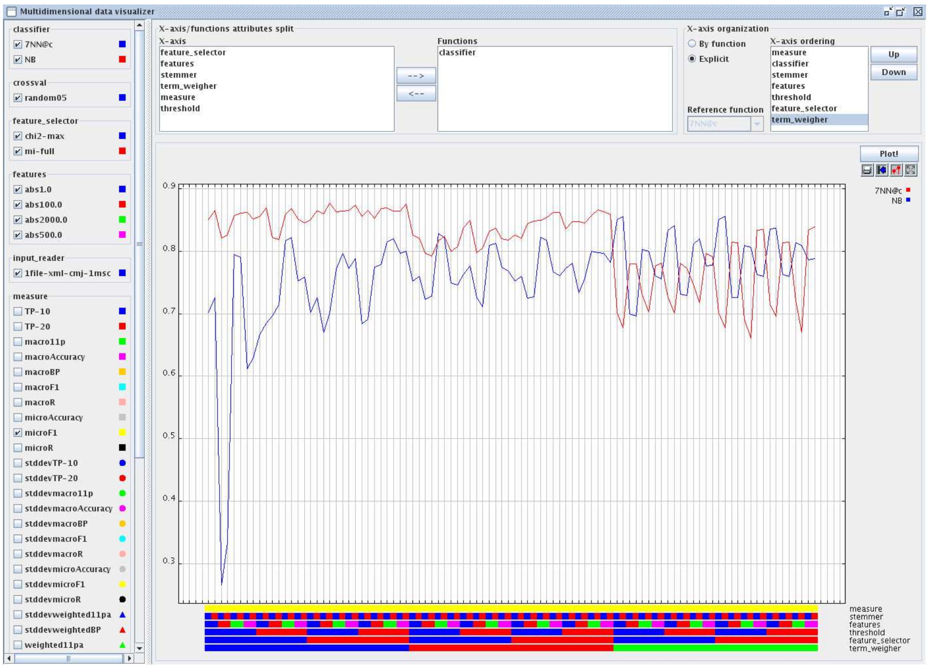
The setup of the experiments is such that we run whole bunch of training attempts in multidimensional learning space of learning methods, features, term weighting types and classifiers:

feature selectors: χ^2, mutual information
feature amount: 100, 500, 2000, all features
term weighting: *bnn*, *nnn*, *atc* [7] (corresponding to binary, term frequency and augmentented TF*IDF weighting schemes in SMART notation)
threshold estimators: fixed, s-cut
classifiers: Naive Bayes, Artificial Neural Network (six hidden units, threshold function tanh), *k*-Nearest Neighbours and Support Vector Machines

For evaluation purposes, we take note on micro/macro F1, TP10, TP20, 11-point-average, accuracy, correlation coefficient, break-even point and their standard deviations for 10-fold crossvalidation.

[10] http://cmj.math.cas.cz/ [11] http://www.emis.de/journals/AM/

All these results are then compared to see which 'points' in the parameter space perform best. This framework allows easy comparison of the evaluated parameters with visualization of the whole result space – see for example multidimensional data visualization on Figure 1. For details see [5].


Multidimensional data visualizer

classifier
☑ 7NN@c 🟦
☑ NB 🟥

crossval
☑ random05 🟦

feature_selector
☑ chi2-max 🟦
☑ mi-full 🟥

features
☑ abs1.0 🟦
☑ abs100.0 🟥
☑ abs2000.0 🟩
☑ abs500.0 🟪

input_reader
☑ 1file-xml-cmj-1msc 🟦

measure
□ TP-10 🟦
□ TP-20 🟥
□ macro11p 🟩
□ macroAccuracy 🟪
□ macroBP 🟧
□ macroF1 🟦
□ macroR 🟥
□ microAccuracy ⬜
☑ microF1 🟨
□ microR ⬛
□ stddevTP-10 ●
□ stddevTP-20 ●
□ stddevmacro11p ●
□ stddevmacroAccuracy ●
□ stddevmacroBP ●
□ stddevmacroF1 ●
□ stddevmacroR ●
□ stddevmicroAccuracy ●
□ stddevmicroF1 ●
□ stddevmicroR ●
□ stddevweighted11pa ▲
□ stddevweightedBP ▲
□ weighted11pa ▲

X-axis/functions attributes split
X-axis
feature_selector
features
stemmer
term_weigher
measure
threshold

Functions
classifier

X-axis organization
○ By function
● Explicit
Reference function
7NN@c

X-axis ordering
measure
classifier
stemmer
features
threshold
feature_selector
term_weigher

Fig. 1. Framework for comparing learning methods. This figure shows comparison of the *k*NN and Naive Bayes classifiers. On the horizontal axis there are particular combinations of the learning space parameters and on the vertical axis the microaveraged F1 measure.

3 Preliminary Results

We know the truth, not only by the reason, but also by the heart. (Blaise Pascal)

Apart from classification, we also tried Latent Semantic Analysis (LSA) [8] to see which concepts are the most relevant.

3.1 Language is relevant

There were papers in several different languages in the CMJ data. After listing the top concepts in LSA of CMJ it is clear that the most significant concepts correspond to language:

1. 0.3*"the" +0.19*"and" +0.19*"is" +0.18*"that" +0.15*"of" +0.14*"we" +0.14*"for" +0.11*"ε" +0.11*"let" +0.11*"then"
2. −0.41*"ist" −0.40*"die" −0.28*"und" −0.26*"der" −0.23*"wir" −0.21*"für" −0.17*"eine" −0.17*"von" −0.14*"mit" −0.13*"dann"
3. −0.31*"de" −0.30*"est" −0.29*"que" −0.27*"la" −0.26*"les" −0.2*"une" −0.2*"pour" −0.20*"et" −0.18*"dans" −0.18*"nous"
4. −0.36*"цхто" −0.29*"для" −0.23*"пусть" −0.19*"из" −0.19*"если" −0.16*"так" −0.16*"то" −0.14*"на" −0.14*"тогда" −0.131169*"мы"
5. −0.33*"semigroup" −0.25*"ideal" −0.19*"group" −0.18*"lattice" +0.18*"solution" +0.16*"equation" −0.16*"ordered" −0.15*"ideals" −0.15*"semigroups" −0.13*"prime"
6. 0.46*"graph" +0.40*"vertices" +0.36*"vertex" +0.23*"graphs" +0.2*"edge" +0.19*"edges" −0.18*"ε" −0.15*"semigroup" −0.13*"ideal" +0.13*"connected"
7. 0.81*"ε" −0.25*"semigroup" −0.16*"ideal" +0.12*"lattice" −0.11*"semigroups" +0.10*"i" −0.1*"ideals" +0.09*"ordered" +0.09*"ř" −0.08*"idempotent"
8. 0.29*"semigroup" −0.22*"space" +0.2*"ε" +0.19*"solution" +0.19*"ideal" +0.18*"equation" +0.16*"oscillatory" −0.15*"spaces" −0.16*"compact" +0.14*"ds"
9. 0.28*"lattice" −0.27*"ε" +0.27*"ordered" +0.23*"group" −0.21*"semigroup" +0.2*"subgroup" −0.19*"ideal" −0.18*"space" +0.16*"groups" +0.16*"torsion"
10. −0.57*"tolerance" −0.22*"compatible" −0.21*"congruence" −0.20*"tolerances" +0.19*"ideal" +0.16*"group" +0.14*"subgroup" +0.13*"prime" −0.13*"algebras" −0.13*"algebra"

First concepts clearly capture the language of the paper (EN, DE, FR, RU), and only then topical itemsets start to be grabbed. It is not surprising – the classifiers then have to be trained either for every language (there is sparsity problem for languages as Czech, Italian or German even French presented in the digital library), or the document features have to be chosen in a language independent manner by mapping words to some common topic ontology. To the best of our knowledge, nothing like EuroWordNet for mathematical subject classification terms or mathematics exists.

3.2 Math notation may be relevant

We also ran LSA on the monolingual corpora of Archivum Mathematicum, where mathematics formulae were not thrown away (recall that this is a subcorpora created from TEX files). Again, taking note of the topmost concepts and their most significant components, we may observe that there appear a few terms containing mathematical formulae (here r and m^n):

1. −0.32*"t" −0.24*"ds" −0.17*"u" −0.17*"_" −0.17*"x" −0.15*"solution" −0.12*"equation" −0.11*"q" −0.11*"x_" −0.11*"oscillatory"

2. 0.28*"ds" +0.28*"t" −0.22*"bundle" −0.16*"natural" +0.15*"oscillatory" −0.15*"vector" +0.13*"solution" −0.13*"connection" −0.13*"manifold" +0.11*"t_0"
3. −0.22*"bundle" +0.19*"ring" −0.17*"natural" −0.16*"oscillatory" +0.15*"fuzzy" −0.15*"ds" +0.12*"ideal" −0.11*"t" −0.11*"r_0" −0.11*"nonoscillatory"
4. 0.29*"ring" −0.23*"x_" −0.21*"_" +0.21*"oscillatory" +0.18*"ideal" +0.17*"r" +0.16*"prime" +0.15*"rings" +0.13*"nonoscillatory" −0.12*"x_n"
5. −0.30*"_" −0.29*"a_" −0.17*"q_" −0.15*"ij" +0.14*"ds" −0.14*"x_" +0.14*"x_n" −0.14*"u_" +0.14*"fuzzy" +0.13*"measurable"
6. 0.87*"fuzzy" +0.19*"x_" +0.10*"oscillatory" +0.10*"ordered" −0.09*"x_n" +0.07*"nonoscillatory" +0.07*"objects" +0.07*"oscillation" −0.06*"ring" −0.06*"periodic"
7. −0.31*"ring" −0.21*"ds" −0.2*"r" −0.17*"rings" −0.17*"ideal" −0.15*"u" +0.13*"oscillatory" −0.12*"prime" +0.12*"curvature" −0.17*"x_"
8. −0.35*"ds" +0.26*"r_0" +0.2*"dx" −0.19*"t_" −0.16*"x_" +0.15*"x_n" −0.15*"holonomic" +0.14*"z" −0.13*"_" +0.12*"natural"
9. −0.24*"r_0" +0.23*"curvature" +0.16*"fuzzy" −0.15*"x_" +0.14*"symmetric" +0.13*"riemannian" +0.13*"m^n" +0.13*"connection" −0.124373*"ordered" −0.124305*"lattice"
10. 0.28*"x_" −0.25*"r_0" −0.24*"ds" −0.18*"fuzzy" +0.15*"oscillatory" +0.14*"holonomic" −0.13*"curvature" −0.12*"u" −0.11*"m^n" +0.1*"oscillation"

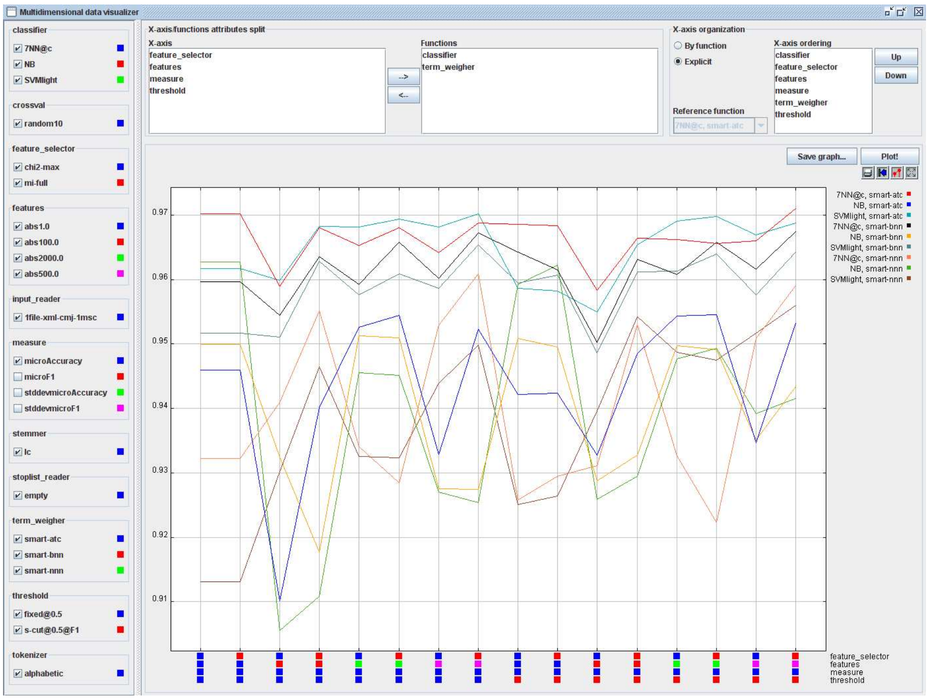
3.3 MSC classification can be learned

Detailed evaluation of classification accuracy shows that with almost all methods we easily reach about 90 % classification accuracy to classify the first two letters of primary MSC. With fine-tuning the best method (Support Vector Machine with Mutual Information feature selection, *atc* term weighting and 500–2000 features) we can increase the accuracy to 95 % or more.

4 Conclusions and Future Work

> Words differently arranged have a different meaning,
> and meanings differently arranged have different effects.
> (Blaise Pascal)

The results presented show feasibility of machine learning approach to the classification of mathematical papers. Given enough data, when we extrapolate the results of preliminary experiments with linear machine methods (creating separable convex spaces in multidimensional feature space) we could approach very high accuracy 98 % or even more. With ambitions for even higher accuracy, higher order models (deep networks) should be used. Mainstream machine learning research was concentrated on using “convex”, shallow methods (SVM, neural networks with backpropagation training) so far. State-of-the-art fine



Multidimensional data visualizer

classifier
☑ 7NN@c 🟦
☑ NB 🟥
☑ SVMlight 🟩

crossval
☑ random10 🟦

feature_selector
☑ chi2-max 🟦
☑ mi-full 🟥

features
☑ abs1.0 🟦
☑ abs100.0 🟥
☑ abs2000.0 🟩
☑ abs500.0 🟪

input_reader
☑ 1file-xml-cmj-1msc 🟦

measure
☑ microAccuracy 🟦
□ microF1 🟥
□ stddevmicroAccuracy 🟩
□ stddevmicroF1 🟪

stemmer
☑ lc 🟦

stoplist_reader
☑ empty 🟦

term_weigher
☑ smart-atc 🟦
☑ smart-bnn 🟥
☑ smart-nnn 🟩

threshold
☑ fixed@0.5 🟦
☑ s-cut@0.5@F1 🟥

tokenizer
☑ alphabetic 🟦

X-axis/functions attributes split

X-axis
feature_selector
features
measure
threshold

Functions
classifier
term_weigher

X-axis organization
○ By function
● Explicit

Reference function
7NN@c, smart-atc

X-axis ordering
classifier
feature_selector
features
measure
term_weigher
threshold

Fig. 2. Side by side comparison of classifier and term weighting performance. Each coloured line represents performance of a combination of classifier (SVM, *k*NN or Naive Bayes) together with a term weighter (*atc*, *bnn* or *nnn*). The evaluation measure here is microaveraged accuracy, that is, the portion of correctly classified test examples. We may see that *k*NN and SVM outperform Naive Bayes and both work consistently best with the *atc* term weighting.

tuned methods allow very high accuracy even on large scale classification problems. However, training of these methods is exceptionally high and the models are big. Using the ensambles of classifiers make the situation even worse (size even bigger), and the final models need to be regularized.

Training large models with non-convex optimization [10] may give classifications that does not exhibit overfitting.

Further studies will encompass fine-grained classification trained on bigger collections, scaling issues, and fine-tuning the best performance by choosing the best set of preprocessing parameters and machine learning methods.

Acknowledgements This study has been partially supported by the grants 1ET208050513 and 1ET208050401 of Academy of Sciences Czech Republic.

References

1. Jackson, A.: The Digital Mathematics Library. Notices of the AMS **50** (2003) 918–923.
2. Bouche, T.: Towards a Digital Mathematics Library? (2006) accepted for publication as a book chapter in Communicating mathematics in the digital era (CMDE) 2006 by A.K. Peters.
3. Sojka, P.: From Scanned Image to Knowledge Sharing. In Tochtermann, K., Maurer, H., eds.: Proceedings of I-KNOW '05: Fifth International Conference on Knowledge Management, Graz, Austria, Know-Center in coop. with Graz Uni, Joanneum Research and Springer Pub. Co. (2005) 664–672.
4. Bouche, T.: A pdfLaTeX-based automated journal production system. TUGboat **27** (2006) 45–50.
5. Pomikálek, J., Řehůřek, R.: The Influence of Preprocessing Parameters on Text Categorization. International Journal of Applied Science, Engineering and Technology **1** (2007) 430–434.
6. Krovetz, R.: Viewing morphology as an inference process. In: Proceedings of the Sixteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Linguistic Analysis (1993) 191–202.
7. Lee, J.H.: Analyses of multiple evidence combination. In: Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Combination Techniques (1997) 267–276.
8. Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer, T.K., Harshman, R.: Indexing by Latent Semantic Analysis. Journal of the American Society for Information Science **41** (1990) 391–407.
9. Ježek, K., Toman, M.: Documents categorization in multilingual environment. In: Proceedings of ElPub 2005, Leuven, Belgium, Peeters Publishing (2005) 97–104.
10. Bengio, Y., Lamblin, P., Popovici, D., Larochelle, H.: Greedy layer-wise training of deep networks. In Schölkopf, B., Platt, J., Hoffman, T., eds.: Advances in Neural Information Processing Systems 19. MIT Press, Cambridge, MA (2007) 153–160.