



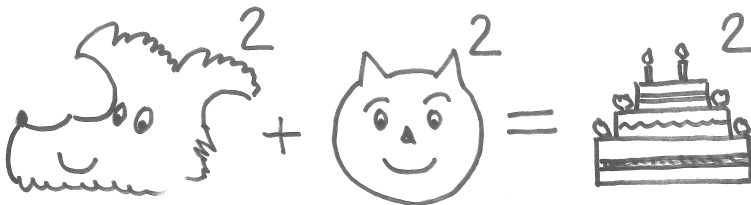
# Indexování a vyhledávání matematických formulí

**Vlastimil Krejčír**, [krejcir@ics.muni.cz](mailto:krejcir@ics.muni.cz)

Inforum 2018, 29.–30. května 2018, Praha



## Kvíz: poznej rovnici



# Motivace

Jak to vlastně začalo?

Rok 2005: **Česká digitální matematická knihovna (DML-CZ).**

Rok 2008: **Evropská digitální matematická knihovna (EuDML).**

Přirozeně vyvstala otázka:

## A co hledání matematických formulí?

- Normální plnotextové hledání na formulích nefunguje.
- V matematice jsou myšlenky vyjádřeny formulemi.
- Matematikům (a příbuzným disciplínám) to může přinést prospěch.
- (Je to zajímavý problém, pojďme se tím zabývat.)

## Motivace II

Q: 'What functionality and incentives would made a working mathematician to login and use a modern DML as EuDML?'

A: 'Math formulae search.'

***Prof. James Davenport,***  
*CEIC member,*  
*MKM 2011 PC chair,*  
*on panel at DML 2011 workshop in Bertinoro as a reply.*

# Jak na to

Hledáme odpovědi na následující otázky:

- Proč matematika nefunguje na „normálním“ vyhledávání?
- Jak to tedy vyřešit?
  - Jak zakódovat matematické formule, aby byly strojově zpracovatelné?
  - Jak matematické formule extrahovat z textů (např. historických skenovaných)?
  - Jak získané matematické formule indexovat a následně porovnávat?
  - Jakým způsobem zapsat formuli jako vyhledávací dotaz?
  - Jak třídit a vážit výsledky vyhledávání?

# Jak funguje běžné vyhledávání

- běžné vyhledávání v textu je „snadné“
- známé algoritmy, dostupný software
- umí to téměř každý informační systém nebo web

# Jak funguje běžné vyhledávání II

## Příklad:

- hledám informace o konferenci Inforum 2018
- do Google zadám text „Inforum 2018“
- Google jej porovnává s obsahem svého indexu
  - který si průběžně buduje procházením webu
  - např. text „Inforum 2018“ je na URL <https://www.inforum.cz/>
- počítač (Google) porovnává způsobem znak po znaku:

I	n	f	o	r	u	m	2	0	1	8	(hledaný dotaz)
											se rovná
I	n	f	o	r	u	m	2	0	1	8	(index Googlu)

# Matematika a textové vyhledávání

Uvedené normální textové porovnávání u matematiky selhává.

Matematika používá abstraktní symboly (proměnné), operace, čísla, ...

**Formule vyjadřující jednu myšlenku lze zapsat různými způsoby,**  
ale stále to bude pro člověka totéž.

Pro počítač to z pohledu textového vyhledávání totéž není!



## Příklad s čísly

$$0,5 = \frac{1}{2} = 2^{-1}$$

$$\sqrt{8} = 2\sqrt{2}$$

Google má v indexu **modrý zápis ve zlomku**, uživatel zadal **dotaz červeně**, porovnáváme znak po znaku:

znak 0  $\neq$  znak 1, znak ,  $\neq$  znak -, ...

Již první znak nesouhlasí! Google nic nenašel, ale měl by...

# Příklad: známé vzorce

## Pythagorova věta

$$a^2 + b^2 = c^2$$

## Příklad: známé vzorce

### Pythagorova věta

$$a^2 + b^2 = c^2$$

je totéž co

$$b^2 + a^2 = c^2$$

## Příklad: známé vzorce

### Pythagorova věta

$$a^2 + b^2 = c^2$$

je totéž co

$$b^2 + a^2 = c^2$$

je totéž co

$$x^2 + y^2 = z^2$$

## Příklad: známé vzorce

### Pythagorova věta

$$a^2 + b^2 = c^2$$

je totéž co

$$b^2 + a^2 = c^2$$

je totéž co

$$x^2 + y^2 = z^2$$

a jsou to speciální případy **Velké Fermatovy věty**

$$a^n + b^n = c^n$$

*Doporučuji knihu:* Simon Singh: Velká Fermatova věta

# Jak zakódovat matematiku: MathML

**MathML:** reprezentace formule v XML, v podstatě se jedná o HTML zápis matematiky (současné prohlížeče MathML znají)

```
<math>
  <mfrac>
    <mn>1</mn>
    <msup>
      <mi mathvariant="bold">x</mi>
      <mn>2</mn>
    </msup>
  </mfrac>
</math>
```

$$\frac{1}{x^2}$$

# Jak získat matematiku ve formě MathML

Jak extrahovat matematické vzorce z existující literatury a zakódovat je do MathML: velmi obtížná úloha...

- tištěné a digitalizované materiály
  - Infty Reader (speciální forma OCR)
- born-digital publikace
  - $\text{L}^{\text{A}}\text{T}_{\text{E}}\text{X}$ ML
- nástroje pracující přímo s MathML
  - MATLAB
- rovnou psát v MathML...

# Jak získané matematické formule indexovat a následně porovnávat?

- formule se rozloží na části (podformule)
- proměnné a čísla (konstanty) se unifikují, tj. nahradí zástupným symbolem

Všechny varianty Pythagorovy věty (např.  $a^2 + b^2 = c^2$ ) nakonec dopadnou přibližně takto:

$$var1^{const1} + var2^{const1} = var3^{const1}$$



## Jak zadávat dotazy?

- matematici jsou uvyklí na systém  $\text{L}^{\text{A}}\text{T}_{\text{E}}\text{X}$ , formuli tedy napíší pomocí známé notace tohoto systému
- píší přímo do vyhledávacího formuláře v prohlížeči (v systému, který matematické hledání podporuje, viz dále)
- formule je převedena pomocí software  $\text{L}^{\text{A}}\text{T}_{\text{E}}\text{X}$ ML do MathML...
- ...a zároveň se při zadávání formule hned vykresluje (systém MathJax)

$a^2 + b^2 = c^2$

Formula preview

$$a^2 + b^2 = c^2$$

Ukázka z webu EuDML (Evropské digitální matematické knihovny)

## Porovnání dotazu a indexu

Formule dotazovaná uživatelem se zpracuje stejným procesem, jakým se formule indexují. Tj. provede se její rozložení a unifikace. Uživatel se dotazuje na Pythagorovu větu  $x^2 + y^2 = z^2$ , ta po zpracování vypadá následovně:

$$var1^{const1} + var2^{const1} = var3^{const1}$$

## Porovnání dotazu a indexu

Formule dotazovaná uživatelem se zpracuje stejným procesem, jakým se formule indexují. Tj. provede se její rozložení a unifikace. Uživatel se dotazuje na Pythagorovu větu  $x^2 + y^2 = z^2$ , ta po zpracování vypadá následovně:

$$var1^{const1} + var2^{const1} = var3^{const1}$$

a v indexu dle formule  $a^2 + b^2 = c^2$  (viz dva slajdy zpět) máme

$$var1^{const1} + var2^{const1} = var3^{const1}$$

## Porovnání dotazu a indexu

Formule dotazovaná uživatelem se zpracuje stejným procesem, jakým se formule indexují. Tj. provede se její rozložení a unifikace. Uživatel se dotazuje na Pythagorovu větu  $x^2 + y^2 = z^2$ , ta po zpracování vypadá následovně:

$$var1^{const1} + var2^{const1} = var3^{const1}$$

a v indexu dle formule  $a^2 + b^2 = c^2$  (viz dva slajdy zpět) máme

$$var1^{const1} + var2^{const1} = var3^{const1}$$

tedy máme **totožné formule!**

# Jak třídit a porovnávat výsledky hledání?

- výsledků hledání je obvykle více
- je nutné je setřídít dle relevance
- kromě unifikovaného zápisu se indexují i původní varianty proměnných a jejich pořadí
- pokud se dotaz shoduje s indexem i v názvech proměnných a jejich pořadí, pak tento záznam má větší relevanci  $\Rightarrow$  dostává se ve výsledcích hledání na vyšší pozice

# MIaS: teorie uvedená do praxe

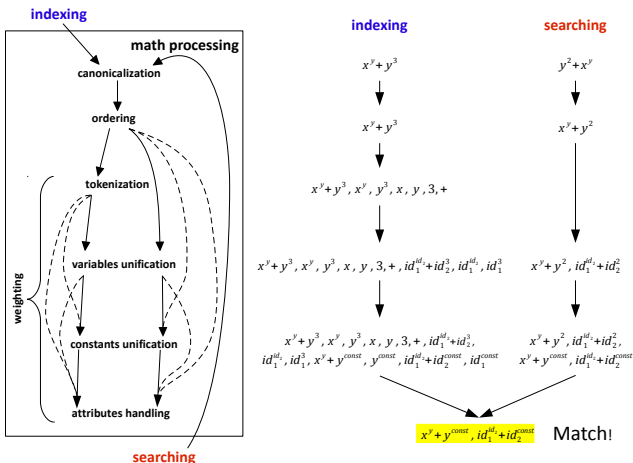
**MIaS** (Math Indexer and Searcher) je volně dostupný open-source software (v jazyce Java) vyvinutý na Fakultě informatiky Masarykovy univerzity, který implementuje předchozí nastíněné techniky a postupy.

## *Technická poznámka pro zvědavé:*

Z technického hlediska je situace komplikovanější. MIaS používá search engine SOLR a unifikované formule kóduje do tzv. *M-Termů*, které již umí SOLR přijmout jako prosté textové řetězce. Příklad *M-termu*:

$$F(N(1)J(I[V=B](1)N(2)))$$

# MIaS: schéma pro zájemce



## Jak to vypadá v praxi

- navržené postupy poměrně dobře fungují
- klíčový problém je získání vstupních dat (zejména pomocí OCR)

### Nasazené systémy:

- **WebMlaS** (webové rozhraní systému MlaS) – používá kvalitní vstupní data z *arXiv.org*
- **EuDML** – Evropská digitální matematická knihovna
- **DML-CZ** (na systému DSpace) – technicky jsme zvládli, není veřejně nasazeno pro nekvalitní vstupní data z OCR



# Příklad WebMIaS

Match any of the following rules

Any field

Add clause

Contains the following formula:

$(S_1 \cup S_2) / S_3$

Rendered:  $\{S_1 \cup S_2\} / S_3$

Search using: presentation and content

Search in: 100-12-100

Verbose output: ☐

Extract subformulae: ☐

Reduce weights of derived formulae: ☐

Search

Total hits: 6852, showing 1-20. Core searching time: 52 ms Total searching time: 82 ms

[math0005151\\_1\\_13.xhtml](#)

...  $(S_1 \cup S_2) / S_3$  ...

score = 18.421957

[xhtml5/L/math0005151/math0005151\\_1\\_13.xhtml.zipmath0005151\\_1\\_13.xhtml](#) - cached XHTML

[math0005151\\_1\\_11.xhtml](#)

... [Math Processing Error] ...

score = 4.966886

[xhtml5/L/math0005151/math0005151\\_1\\_11.xhtml.zipmath0005151\\_1\\_11.xhtml](#) - cached XHTML

[0806.4024\\_1\\_146.xhtml](#)

... has a small number of isometry types with respect to the action on  $\mathcal{X} / \approx$  ...

score = 0.09666692

[xhtml5/5/0806.4024/0806.4024\\_1\\_146.xhtml.zip0806.4024\\_1\\_146.xhtml](#) - cached XHTML

[1108.5123\\_1\\_63.xhtml](#)

... and let  $E / \approx$  be the metric quotient. ... is second countable and locally compact then  $E / \approx$  is second countable and locally compact too.

score = 0.06792955

[xhtml5/1108.5123/1108.5123\\_1\\_63.xhtml.zip1108.5123\\_1\\_63.xhtml](#) - cached XHTML

[1203.1283\\_1\\_46.xhtml](#)

... l'ensemble quotient  $(\mathbb{R}_+ \times \mathbb{R}_+) / \approx$  ...

score = 0.06045454

Data (matematické formule) vzaty z arXiv.org

# Jiný příklad WebMIaS

Match  of the following rules

Any field

Add clause

Contains the following formula:

$$\phi_H \circ \pi_k \circ \pi \circ M \in C(\bar{X}, S^1)$$

Rendered:  $\phi_H \circ \pi_k \circ \pi \circ M \in C(\bar{X}, S^1)$

Search using:

Search in:

Verbose output: ☐

Extract subformulae: ☐

Reduce weights of derived formulae: ☐

**Search**

Total hits: 4888196, showing 1–20. Core searching time: 15162 ms Total searching time: 15797 ms

[math0005151\\_1\\_101.xhtml](#)

...  $\phi_k = P_{k+1} \in C(\bar{X}, S^1)$  ... where  $\phi_k \in$  ...

score = 1.0956602

[xhtml5/math0005151/math0005151\\_1\\_101.xhtml.zipmath0005151\\_1\\_101.xhtml](#) - cached XHTML

[1105.2779\\_1\\_144.xhtml](#)

... we have  $\phi_k \in P_k \in \mathcal{D}(\mathcal{E})$ , moreover ... then  $(\phi_k \circ P_k)_n$  converges weakly to ...

score = 0.05874735

[xhtml5/1105.2779/1105.2779\\_1\\_144.xhtml.zip1105.2779\\_1\\_144.xhtml](#) - cached XHTML

[math0404322\\_1\\_59.xhtml](#)

... and  $\phi_k \circ g_k \in N(g_k), \dots l. (g_k, \phi_k \circ g_k)$  is ...

score = 0.05278814

[xhtml5/math0404322/math0404322\\_1\\_59.xhtml.zipmath0404322\\_1\\_59.xhtml](#) - cached XHTML

[math0404322\\_1\\_56.xhtml](#)

... then  $\zeta_n \notin K_n$  implies ... , hence  $\phi_n \circ g_n \in N(g_n)$  by ( ...

score = 0.037092865

[xhtml5/math0404322/math0404322\\_1\\_56.xhtml.zipmath0404322\\_1\\_56.xhtml](#) - cached XHTML

[1302.2341\\_1\\_101.xhtml](#)

Next we prove that  $\alpha \circ \psi \in AC$ .

score = 0.02955749

Data (matematické formule) vzaty z arXiv.org

# Vývojový tým MIR

Problematiku matematického indexování a vyhledávání řeší výzkumný tým MIR (Maths Information Retrieval) na Fakultě informatiky Masarykovy univerzity:



**Petr Sojka**



**Martin Líška**



**Michal Růžička**

## Odkazy

Domovská stránka týmu MIR:

<https://mir.fi.muni.cz/>

WebMlaS:

<https://mir.fi.muni.cz/webmias-demo/>

EuDML (rozšířené vyhledávání matematiky):

<https://eudml.org/search>

**Máte-li otázky, sem s nimi :-)**

# Kvíz II: „pejsek a kočička“ v praxi

Quiz 4

Name

1. (15 pts) Find the general solution of the homogeneous differential equation.



$$\frac{dy}{dx} + 4\frac{dy}{dx} + 3y = 0$$

*+15*

$$y'' + 4y' + 3y = 0$$

Assume  $y = e^{rt}$



$$r^2 + 4r + 3 = 0$$

LET  $r =$    $r =$  

$$(r + 3)(r + 1) = 0$$

*Technically it's not never again!!*

$$r = -3, -1$$

So...  $C_1 =$    $C_2 =$  

$$y = e^{-3x} + e^{-x}$$

*Necessary?*

*HAPPY DOG NOT HAPPY CAT*

*WTF??*