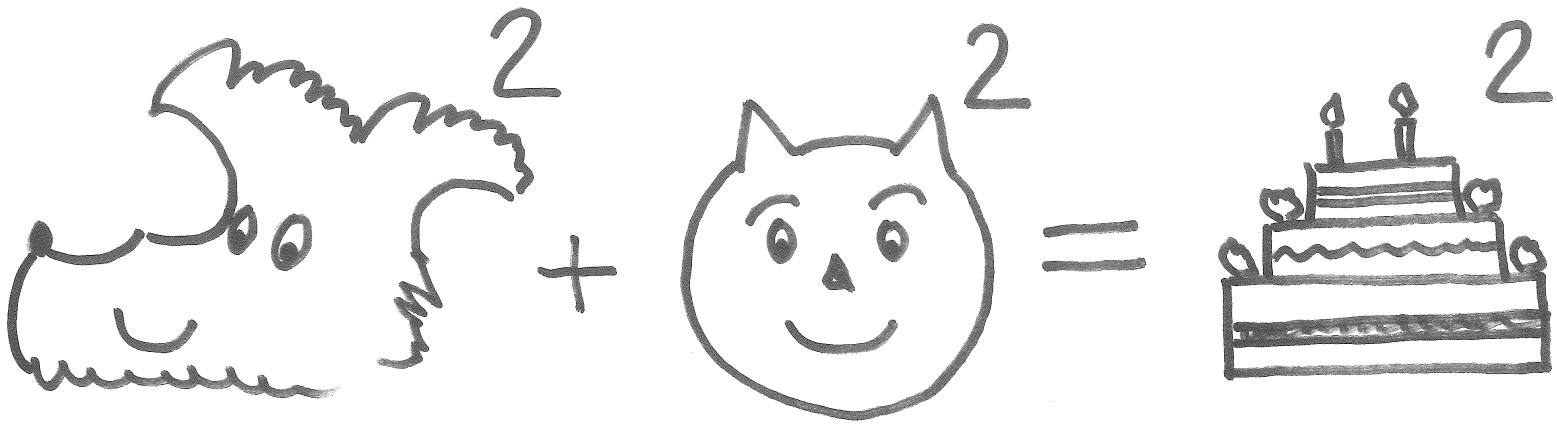
ÚSTAV VÝPOČETNÍ TECHNIKY
Masarykova univerzita

Indexování a vyhledávání matematických formulí

Vlastimil Krejčíř, `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 Frematovy věty**

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

Doporučuji knihu: Simon Singh: Velká Frematova 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>
```

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

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
  - LaTeXML
- 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 LaTeX, 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 LaTeXML do MathML...
- ...a zároveň se při zadávání formule hned vykresluje (systém MathJax)

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 setřídit 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

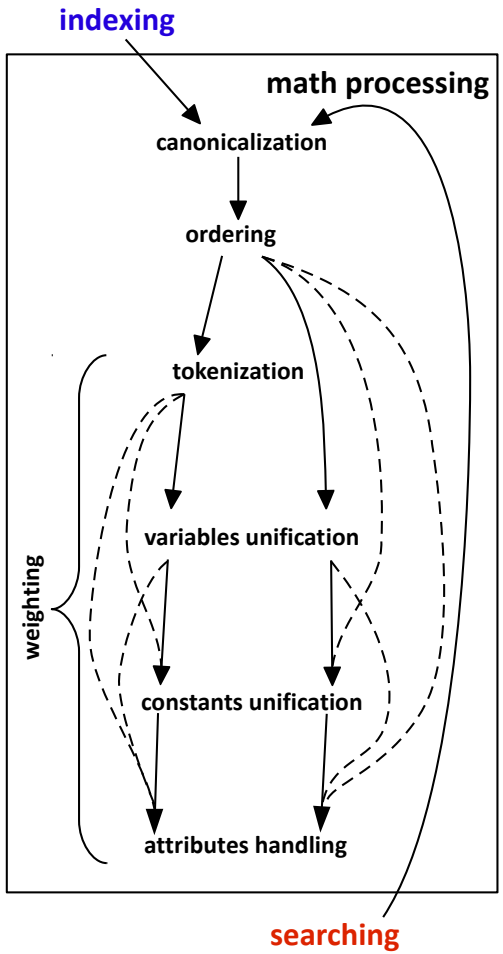
MIaS (Math Indexer and Searcher) je volně dostupný open-source software (v jazyce Java) vyvinutý na Fakultě informatiky Masarykovy univerzity, který implementuje předchozí nastíněné techniky a postupy.

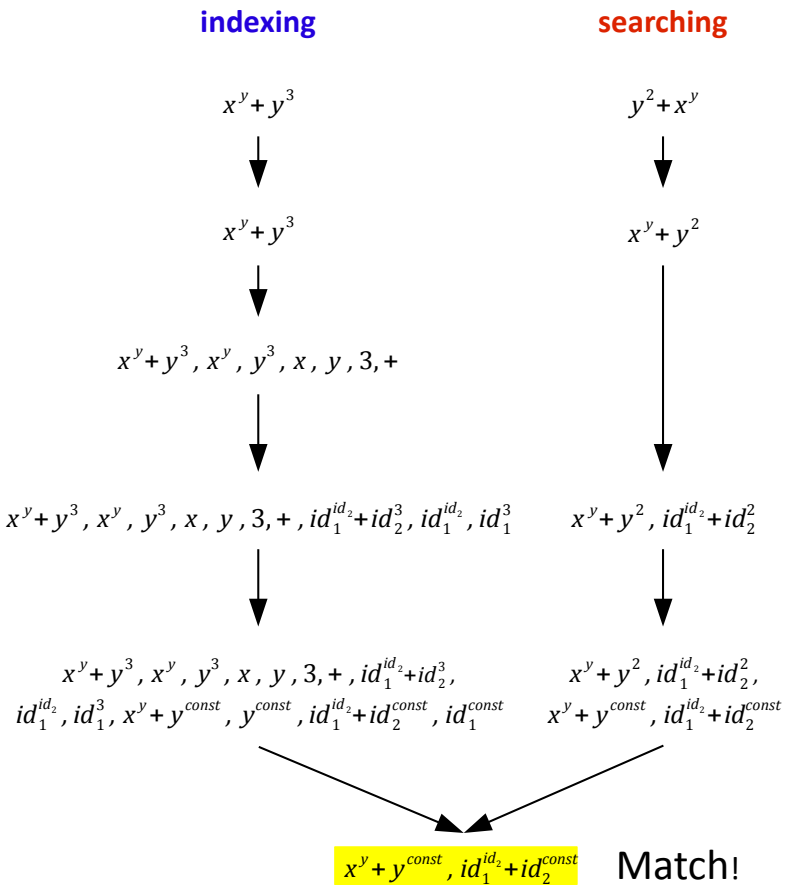
Technická poznámka pro zvědavé:

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:

- **WebMIaS** (webové rozhraní systému MIaS) – 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

Data (matematické formule) vzaty z arXiv.org

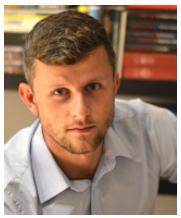
Jiný příklad WebMIaS

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/`

WebMIaS:

`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