# Indexování a vyhledávání matematických formulí

**Vlastimil KREJČÍŘ**
Ústav výpočetní techniky, Masarykova univerzita, Brno
krejcir@ics.muni.cz

INFORUM 2018: 24. ročník konference o profesionálních informačních zdrojích
Praha, 29. – 30. 5. 2018

***Abstrakt***

Vyhledávání v digitálních textových dokumentech je dnes poměrně dobře zvládnutou disciplínou a běžnou součástí mnoha informačních systémů. Existuje však řada speciálních aplikací, kde běžné známé metody indexace a plnotextového vyhledávání selhávají. Článek se populární formou věnuje problému indexace a vyhledávání matematických vzorců, které jsou důležitou součástí odborných dokumentů. Na řadě příkladů je uvedena celá problematika a způsob, jakým ji lze řešit.

---

Indexace a vyhledávání v textech je dnes v digitálním světě dobře zvládnutou technologií a součástí mnoha informačních systémů. Uživatelé považují možnost vyhledávat v plných textech – mít ono univerzální okénko pro vložení dotazu jako je tomu ve vyhledávači Google – za naprostou samozřejmost. Stále se však pohybujeme v oblasti hledání běžné textové informace, kdy dotaz snadno napíšeme na klávesnici počítače. Existuje však řada speciálních aplikací, kde známé metody indexace a plnotextového hledání selhávají. Typickým příkladem jsou obory, které pracují s matematickými formulemi: samotná matematika, fyzika, architektura, strojírenství, …

Potřebujeme však vyhledávat v matematických formulích? Je to pro vědce skutečně přínosné? Motivací lze najít více – primární je však zájem odborné komunity, který vyvstal v minulých letech, kdy docházelo k rozvoji digitálních knihoven specializovaných na matematickou odbornou literaturu. Primárně byly tyto knihovny budovány na národní úrovni, později vznikl celoevropský projekt EuDML, který zapojil a agregoval obsah národních matematických digitálních knihoven a umožnil je prohledávat všechny z jednoho místa.

Vznik matematických knihoven často iniciovali sami matematici a na jejich tvorbě se podíleli. Možnost speciálního prohledávání odborných textů obsažených v těchto knihovnách se i jim jevila a jeví jako zajímavé a přínosné rozšíření, citujme:

*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.*

V České republice je možné pracovat s Českou digitální matematickou knihovnou (DML-CZ)[1], shromažďující odbornou matematickou literaturu, která kdy vyšla na historickém území českých zemí. V současné době ji provozuje Matematický ústav Akademie věd ČR. Na jejím vývoji spolupracovala řada partnerů v ČR, mj. i vědci z Fakulty informatiky Masarykovy univerzity, kteří

1 Česká digitální matematická knihovna (DML-CZ), dostupná na https://dml.cz/

se pod vedením doc. Petra Sojky začali zabývat problémem indexování a vyhledávání matematických formulí. Cílem bylo nalézt řešení, které by bylo možné uvést do praxe – přesné a pro koncové uživatele dostatečně komfortní vyhledávání matematických formulí. Tedy umožnit uživatelům pomocí speciálního formuláře ve webovém prohlížeči zadat matematickou formuli (například „$a^2 + b^2 = c^2$“) a jako výsledek vrátit odkazy na články, ve kterých je tato formule použita. Řešení tohoto úkolu iniciovalo vznik výzkumné skupina **MIR** (*Maths Information Retrieval*), která se problému intenzivně věnovala a stále věnuje.

## *Jak na vyhledávání z pohledu počítače*

Abychom lépe pochopili, jaké těžkosti s sebou práce s matematickými formulemi z pohledu počítače nese, nastíníme si zjednodušeně, jak pracuje počítač při hledání v běžném textu. Pro lepší představu použijeme příklad s vyhledávačem Google. Hledáme-li například informace o konferenci Inforum 2018, pak do vyhledávače Google zadáme prostý dotaz „Inforum 2018“ a Google nám okamžitě vrací výsledky seřazené dle relevance (tak jak ji určil dle svých pravidel on sám).

Z pohledu počítačového je dotaz „Inforum 2018“ pouze zakódovaným řetězcem znaků, tedy „I“, „n“, „f“, … Po odeslání dotazu do vyhledávače Google je tento řetězec na tyto jednotlivé znaky rozložen, a jednotlivě znak po znaku (tedy „I“ hledané = „I“ v indexu, „n“ = „n“, …) porovnáván s obsahem indexu (databáze obsahu webu, kterou si Google průběžně buduje a aktualizuje), ve které má tento řetězec Google indexován spolu s informací, na které webové stránce tento řetězec při indexaci nalezl. Odkaz na nalezenou stránku pak dá jako výsledek hledání. Tedy hledaný výraz „Inforum 2018“ souhlasí s indexovaným „Inforum 2018“ a výsledek hledání je jednoznačný.

Uvažujeme-li matematické formule, výše uvedený jednoduchý způsob porovnávání nemůže správně fungovat. Důvodem je syntaktická (nikoli sémantická!) nejednoznačnost zápisu v matematice. Vezměme příklad *Pythagorovy věty*, obvyklý zápis

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

je z matematického hlediska totéž co

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

Pokud by uživatel zadal do vyhledávače první variantu Pythagorovy věty a vyhledávač měl klasickým způsobem indexovanou druhou variantu, pak výsledek bude, že daná formule nebyla nalezena. Vyhledávač totiž tyto formule neztotožní – začne je porovnávat znak po znaku a shodu vyloučí hned u znaku prvního, protože „a“ je jiný znak než „x“. Proto je třeba k matematickému vyhledávání a indexování přistupovat zcela jiným způsobem.

Dále v článku se podrobněji podíváme, jak se s podobnými problémy, spojenými s počítačovým zpracováním matematiky, vypořádat. Klíčové otázky, které při důkladnějším rozboru celé problematiky vyplynou, jsou následující:

- 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 napsat vyhledávací dotaz?
- Jak třídit a vážit výsledky vyhledávání?

## *Jak zakódovat matematické formule?*

Odpovědí na otázku je jazyk **MathML** (*Mathematical Markup Language*), standard konsorcia W3C, které je dnes zodpovědné na standardizaci v oblasti webu (hlavně jazyka HTML). MathML je v podstatě specializovaná varianta jazyka XML a slouží pro zápis matematických formulí. Například kód

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

se bude vykreslovat (renderovat) na zlomek

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

Jazyk MathML je dnes součástí všech běžně rozšířených prohlížečů. To znamená, že pokud výše uvedený kód vložíte do kódu HTML, pak jej prohlížeče rozpoznají a zobrazí daný matematický výraz. MathML se již v praxi hojně používá, viz například výklad Pythagorovy věty na Wikipedii[2].

## *Jak získat matematické formule v MathML?*

Abychom mohli matematické formule indexovat, potřebujeme je mít zakódované v MathML. Jejich získání je velmi nesnadný úkol, ať již uvažujeme odborné texty dostupné pouze v tištěné podobě, nebo i v podobě digitální. U historických tištěných textů je nutné provést specializovanou formu OCR[3], schopnou na naskenované stránce rozeznat matematickou formuli, v ní jednotlivé znaky a vše převést do MathML. V současné době tento problém není dostatečně vyřešen – existuje varianta programu Infty Reader, která OCR matematických formulí umí, výsledky jsou bohužel velmi neuspokojivé a pro praktické nasazení špatně použitelné.

Varianta, kdy již máme odborný text v digitální podobě, je zajímavější. Matematická komunita s oblibou používá pro sazbu matematických textů systém TeX (a jeho varianty LaTeX, AmsTeX apod.). Bohužel systém TeX dává autorům poměrně dosti volnosti v podobě tvorby vlastních maker a to strojový převod do MathML velmi komplikuje. I přesto se jedná o perspektivní způsob, jakým MathML získat, a proto pro konverzi ze systému TeX do MathML existuje řada programů (LaTeXML).

K dispozici jsou také programy pro matematickou komunitu, které s jazykem MathML přímo pracují a jejichž výstupem mohou formule v MathML být. Jmenujme alespoň rozšířený software pro matematické výpočty MATLAB.

---

2 Pythagorova věta na české Wikipedii: https://cs.wikipedia.org/wiki/Pythagorova_v%C4%9Bta.

3 OCR – Optical Charecter Recognition: metoda digitalizace skonovaných textů, viz https://cs.wikipedia.org/wiki/OCR.

## *Jakým způsobem zadávat dotazy?*

Z hlediska zpracování počítačem by optimální bylo, kdyby uživatelé zadávali dotazy na matematické formule přímo v jazyce MathML. Ten je ale uživatelsky málo přívětivý. Matematická odborná komunita je uvyklá použití notace systému TeX. Proto existuje software, který umožňuje přímo v prohlížeči formuli zapsanou v notaci TeXu vykreslovat (renderovat) za běhu (on-the-fly). To zároveň slouží zadavateli i jako zpětná kontrola, píše-li formuli správně.

## *Jak matematické formule indexovat?*

Jedná se v celém procesu práce s matematickými formulemi o klíčovou otázku. Problémy spojené s indexací matematických formulí jsme již mírně nastínili v předešlém textu, podívejme se proto na ně nyní podrobněji.

Kromě již uvedené záměny proměnných (tedy že z matematického hlediska je $a^2 + b^2 = c^2$ totéž co $x^2 + y^2 = z^2$), lze objevit i řadu dalších pro počítačové zpracování problémových míst. Například když ve vzorci Pythagorovy věty ponecháme proměnné $a$, $b$ a $c$, ale prohodíme jejich pořadí

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

To je opět Pythagorova věta, pouze jsme využili toho, že při sčítání nezáleží na pořadí sčítanců. Můžeme i uvažovat vzorec

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

jedná se o *Velkou Frematovu větu*[4], jež je zobecněním věty Pythagorovy.

Ale lze nalézt i mnohem triviálnější problémy, které komplikují strojové porovnávání formulí. Bude nám stačit jedno číslo, číslo reprezentující jednu polovinu. Platí následující rovnost:

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

Z hlediska počítačového porovnávání dle znaků jde o zcela různé řetězce znaků. Komplikovanější příklad z oblasti úpravy odmocnin:

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

Tým MIR z Fakulty informatiky navrhl a implementoval postupy, jak se s výše uvedenými problémy vypořádat. Jsou založeny na principu rozložení a zobecnění matematické formule. Celý postup při indexaci lze shrnout do několika kroků:

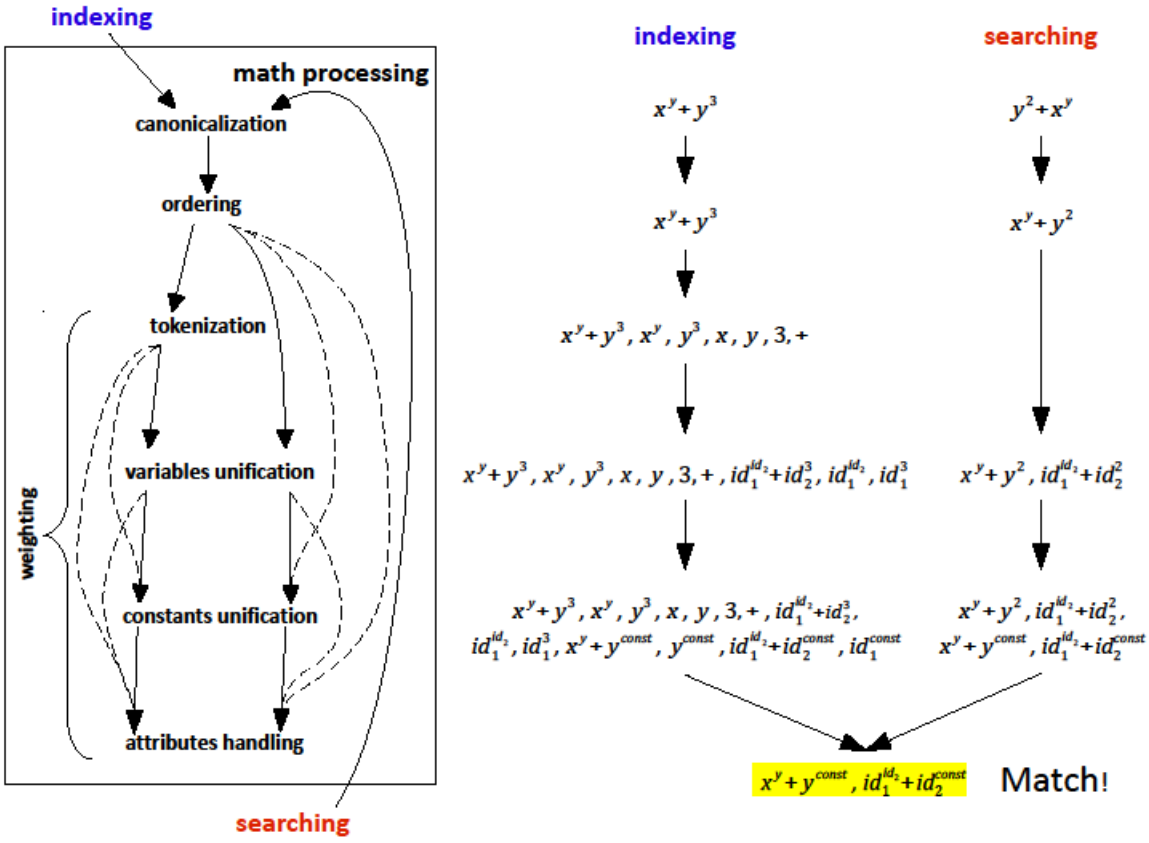
---

4 Velká Fermatova věta říká, že neexistuje žádné přirozené číslo $n$ větší než 2, pro které by rovnice $a^n + b^n = c^n$ měla řešení. Geniální matematik Pierre de Fermat si v 17. století k této rovnici poznačil na okraj jedné z knih, že zná jednoduchý důkaz výše uvedené věty, ale okraj na stránce dané publikace je příliš malý, aby jej tam mohl celý vepsat. Z Fermatovy strany se jednalo pravděpodobně pouze o žert, protože důkaz Velké Fermatovy věty provedl až v roce 1994 americký matematik Andrew Wiles s pomocí prostředků, které Fermat nemohl ve své době znát. Wilesův důkaz patří k jedněm z nejsložitější důkazů v historii matematiky.

- *Canonicalization*: úprava vstupního MathML.
- *Ordering*: seřazení proměnných (například u komutativních operací typu sčítání a násobení).
- *Tokenization*: rozložení formule na části, což následně umožní hledat i podformule, tedy se například dotázat jen $a^2 + b^2$ a systém jako výsledek nabídne články s celou Pythagorovou větou.
- *Unification*: nahrazení proměnných a čísel (konstant) zástupnými symboly.

Když uživatel zadá jako dotaz matematickou formuli, pak s touto formulí proběhne stejný proces, a výsledný kód se porovnává s obsahem indexu. Protože jsou jak indexovaná, tak hledaná formule dostatečně zobecněny, může systém najít shodu i v případě, že se syntakticky dané formule neshodují: systém tak vyhodnotí $a^2 + b^2 = c^2$ a $x^2 + y = z^2$ jako formule totožné. Po zobecnění totiž obě vypadají přibližně takto: $var1^{const1} + var2^{const1} = var3^{const1}$.

Přesnější schéma celého procesu uvádíme v následujícím schématu:




Matematiky znalý čtenář si jistě dokáže představit i mnohem komplikovanější situace, které mohou nastat. Absolutně přesný systém indexace a vyhledávání by musel s matematickými formulemi sémanticky pracovat mnohem hlouběji.

## *Jak třídit a vážit výsledky vyhledávání?*

Při hledání shody v indexu je často nalezena shoda méně přesná, nebo jen shoda s částí formule apod. Zároveň například přirozeně očekáváme, že pokud se hledaná formule shoduje i v názvech a pořadí proměnných a hodnotách konstant, pak bychom takový výsledek chtěli vidět na předních místech. Naopak, čím více se formule syntakticky liší, tím je shoda méně přesná a výsledek hledání by měl být řazen níže. Opět si vezměme příklad Pythagorovy věty a dotaz „$a^2 + b^2 = c^2$“.

Přirozeně očekáváme, že články, ve kterých se věta vyskytuje právě v tomto tvaru, by měly být řazeny ve výsledcích dříve než články, ve kterých se formule vyskytuje ve tvaru s proměnnými $x$, $y$ a $z$. Výsledky vyhledávání je tedy třeba velmi pečlivě vážit a třídit – a opět se jedná o nesnadnou úlohu.

## *Systém MIaS*

Tým MIR výše uvedené postupy implementovat do funkčního software, který je nazván **MIaS** (*Mathematical Indexer and Searcher*). Jedná o open-source knihovnu v jazyce Java, kterou může kdokoli volně použít.

Systém MIaS provede příslušné transformace matematické formule a výsledkem těchto transformací je tzv. *M-term*, matematická formule zakódovaná do obyčejného textového řetězce znaků, který je dále možné indexovat běžnými metodami užívanými pro text. Pro představu uveďme příklad M-termu: `F(N(1)J(I[V=B](1)N(2))).`

V praxi byl systém testován nad daty z repozitáře Arxiv.org a nasazen v Evropské digitální matematické knihovně EuDML, kde si jej může zájemce snadno vyzkoušet. Další možností, jak si hledání formulí vyzkoušet, je webové rozhraní systému MIaS zvané WebMIaS[5]. Testování systému je komplikované – v reálném provozu je třeba, aby uživatelé, kterými jsou v tomto případě odborníci-matematici, sami řekli, je-li vyhledávání dostatečně přesné a funkční.

V současnosti se jako největší problém jeví extrakce matematických formulí z článků a jejich konverze do formátu MathML. Samotné algoritmy systému MIaS fungují poměrně dobře a dostatečně přesně, jejich nasazení brání nekvalitní data pro indexaci. Na Ústavu výpočetní techniky MU jsme provedli integraci systému MIaS do systému Dspace, který slouží jako platforma pro provoz České digitální matematické knihovny DML-CZ. Problém nedostatečně kvalitních vstupních dat (extrakce a převod matematických formulí z článků v DML-CZ) však dosud brání zpřístupnění této funkcionality pro koncové uživatele.

## *Závěr*

Popsali jsme motivaci pro matematické vyhledávání, způsoby, jak jej řešit, i praktickou implementaci. V dané oblasti je ještě řada otevřených problémů a celý systém indexace a vyhledávání matematických formulí je možné dále zkoumat a vylepšovat. Závěrem ještě jmenujme členy MIR týmu Fakulty informatiky Masarykovy univerzity, kteří systém MIaS budují:

doc. RNDr. Petr Sojka, Csc.
RNDr. Martin Líška
RNDr. Michal Růžička, Ph.D.

Zájemci o podrobnější informace mohou navštívit domovskou stránku týmu MIR na adrese https://mir.fi.muni.cz/webmias-demo/, na které naleznou především odkazy na odborné články, které celou problematiku přesně a do hloubky popisují.

5 WebMIaS, webové rozhraní pro testování systému MIaS: https://mir.fi.muni.cz/webmias-demo/