Automatické rozpoznávání a indexování knižních obsahů

Název dokumentu u vědecko-technických monografií nepřináší informace o všech obsazených tématech. Napr. z názvu knihy „XML technologie: Principy a aplikace v praxi" nepoznáme, jestli se v ní píše o jazyce XQuery. V techto případech selhává i vecný popis bibliografického záznamu, protože katalogizátor často nezná do hloubky popisovanou problematiku, navíc omezený počet a míra jemnosti predmetových hesel jsou pro tyto prípady nedostatečné. Naopak, obsah knihy (TOC, table of content) u vědecko-technických dokumentu velmi presne popisuje obsahy jednotlivých kapitol a podkapitol, takže z nich můžeme úspešne dolovat klíčová slova s vysokou relevancí. Taková klíčová slova se pak mohou stát vstupnîmi daty pro indexování ve vyhledávacích nástrojích typu OPAC nebo discovery systém, kde uživatel získá možnost vyhledávat i podle slov a frází vyskytujících se v obsahu díla.

Tištěný knižní fond je v katalozích knihoven tradičně zpracováván v podobě bibliografických záznamů, tvořených jmennými a věcnými údaji. Vybrané údaje jsou v katalogu indexovány a lze podle nich vyhledávat či filtrovat (např. v podobě facetových filtrů). Pokud uživatel vyhledává dokumenty podle tématu, může použít buď názvové údaje nebo údaje věcného popisu (předmětových hesel, klíčových slov, znaků klasifikačních systémů apod.).

Název díla u odborných publikací se sice snaží výstižně pojmenovat celkové téma nebo zaměření (tj. o čem kniha je), nicméně není schopen pojmout všechna témata a podtémata v díle obsažená. Např. název knihy „Vybrané kapitoly z jaderné fyziky" sice naznačuje, že zastřešujícím tématem knihy je jaderná fyzika, netušíme však, o jakých konkrétních tématech se v knize dočteme. V takovém případě jsme odkázáni na věcné údaje, pokud jimi je bibliografický záznam opatřen.

Věcné údaje tradičně vytváří knihovník-katalogizátor s knihou v ruce. Kvalita popisu závisí na intelektuálním úsudku katalogizátora a je vždy zatížena určitou mírou subjektivity. U odborné vědecko-technické literatury dochází logicky k tomu, že knihovník, který není expertem na danou vědní disciplínu, není schopen témata obsažená v dané monografii správně rozlišit a relevantně popsat. Dochází nejen k nepřesnostem popisu, ale často i ke zcela chybnému výběru věcných selekčních údajů, kdy knihovník použije jiný termín, než by použil odborník. Z časových důvodů také nemá knihovník prostor k tomu, aby knihu podrobně prošel a obsažená témata podrobně nasnímal. Koncový uživatel v důsledku není schopen najít v katalogu všechny monografie, které objektivně odpovídají jeho dotazu.

Příkladem může být situace, kdy uživatel hledá monografii, ve které by se dočetl, jakým způsobem může pracovat se souborovým systémem v programovacím jazyce Python. Pokud bude v katalogu hledat podle názvu či podle věcných údajů klíčové slovo „souborový systém AND jazyk python", odpovídající záznamy nenalezne. Důvod je jednoduchý: katalogizátor při tvorbě záznamů pravděpodobně správně usoudí, že příslušná kniha pojednává o programovací jazyce Python, ale už není schopen popsat, že se v některé kapitole knihy píše o práci se souborovým systémem. Za prvé manuálně vytvářený věcný popis nejde do takové hloubky a za druhé katalogizátor není pravděpodobně odborník na programovací jazyky a není schopen takové téma v knize rozpoznat nebo ho ohodnotit jako významné.

V důsledku toho je vyhledávání podle názvových a věcných údajů odborných publikací zatížené významnou chybovostí. Odborný uživatel knihovny často nedokáže v katalogu knihovny objevit díla, která by ke své práci potřeboval. K tomuto problému se mohou knihovny postavit několika způsoby.

Prvním způsobem, bohužel v našich knihovnách nejčastějším, je ignorace tohoto problému, kdy věcný popis vytvářejí napříč celým fondem univerzální katalogizátoři a nepřesnosti popisu se jednoduše tolerují.

Druhým způsobem, jak lze problém řešit, je najmutí věcných katalogizátorů, kteří jsou odborníky na určitou vědní disciplínu, např. biologii. Takto zaměřený katalogizátor je schopen poměrně přesně vytvářet věcný popis knižního fondu z oblasti biologických věd. Toto řešení je však možné pouze v úzce specializovaných odborných knihovnách, kde předmětové rozpětí celého fondu je tak malé, že může knihovna zaměstnat odpovídající množství oborových katalogizátorů. Univerzální knihovny, jako jsou např. krajské vědecké knihovny, však budují fond s tak velkým počtem vědních disciplín, že si nemohou dovolit zaměstnat pro každý obor speciálně zaměřeného katalogizátora.

V takových případech je nutné se zamyslet nad možnostmi vytváření relevantního věcného popisu bez účasti oborových katalogizátorů. Jednou z možností je automatické rozpoznávání knižních obsahů s vytěžováním relevantních klíčových slov.

Klíčová slova obsažená v obsahu monografií

Odborné monografie jsou ve většině případů opatřeny informačním aparátem, který čtenáři umožňuje lépe se orientovat v textu a rychle přecházet do hledaných kapitol či na konkrétní stránku. Nejčastěji se jedná o abecední rejstřík knihy a o obsah knihy.

Zatímco knižní rejstřík obsahuje značně atomizované a izolované výrazy, které jsou příliš závislé na kontextu, knižní obsah je ideální pro získání představy o tématech, o kterých odborná monografie pojednává. Obsah knihy tvoří názvy a podnázvy částí dokumentu, typicky kapitol, v pořadí, ve kterém se v knize vyskytují, s označením stránky, kde příslušné části začínají. Názvy kapitol vytváří sám autor, který text dokonale zná, a měl by proto být schopen názvem přesně vystihnout obsah kapitoly. U odborného textu tak názvy kapitol tvoří klíčová slova, pečlivě volená autorem, normalizovaná do základních tvarů, s důrazem na stručnost a výstižnost. Jednoduše řečeno, název kapitoly odborného textu vyjadřuje obsah kapitoly. To obecně neplatí

u beletrie, poezie a další umělecké literatury, kde jsou názvy kapitol či básní voleny s jiným záměrem, než pouze stroze vyjádřit obsah.

Knižní obsah odborných monografií se tak stává ideálním zdrojem pro vytěžování klíčových slov, které můžeme následně využít k obohacení věcného popisu monografií. Mohlo by se zdát, že tato metoda je podobná získávání klíčových slov z plných textů dokumentu, ale není tomu tak. Základní výhody této metody, oproti metodám vytěžování klíčových slov z plného textu dokumentu, můžeme spatřit v následujících aspektech:

1. nemusíme mít k dispozici plný text, stačí nám získat pouze několik stran s obsahem,
2. položky knižního obsahu jsou prezentovány v základních mluvnických tvarech,
3. nejsou závislé na kontextu nebo je kontext jasně zřejmý díky hierarchickým vztahům položek (název díla —> název oddílu —> název kapitoly —> název podkapitoly apod.),
4. díky odkazům na stránky, kde příslušné kapitoly začínají, můžeme určovat také váhu klíčových slov podle toho, na kolika stránkách se o daném tématu píše.

Výsledkem této metody může být sada klíčových slov opatřených parametrem, který vyjadřuje prostor, jenž autor danému tématu v knize věnoval. To může být využito při vyhledávání k řazení výsledků podle relevance.

Metoda

Podívejme se nyní podrobněji, jakým způsobem může být metoda vytěžování klíčových slov z knižních obsahů prakticky realizována na názorném příkladu. Předpokládejme, že budeme zpracovávat monografie, které máme pouze v tištěné podobě. Zpracování obsahu rozdělme do následujících kroků:

1. Naskenování stránek knihy, na kterých se vyskytuje obsah
2. OCR s rozpoznáním bloků textu
3. Eliminace cizích bloků
4. Rozlišení textových a číselných bloků, odstranění stopslov
5. textová analýza s důrazem na kontext položek obsahu
6. Získání klíčových slov
7. Využití

Naskenování stránek knihy, na kterých se vyskytuje obsah

Z knihy skenujeme všechny stránky, kde je obsažen obsah knihy. Podle toho, jak je obsah podrobný, je rozložen na jedné stránce nebo i na desítkách stran. Často se setkáme s tím, že je na stránkách s obsahem přítomen i jiný text, případně i nějaké grafické objekty. Takové cizorodé elementy budou odstraněny v dalších krocích. U skenování obsahu musíme mít na paměti, že primárním účelem je následné získání textu pomocí OCR, je proto nutné kontrolovat vstupní chyby. Problémy může způsobit chybějící stránka, ohnutý roh, křivý tisk na papíře nebo křivě položená kniha na skeneru. Pečlivost snižuje míru chyb v následných krocích. Výsledkem skenování je bitmapový objekt v podobě grafického souboru, např. TIFF.

OCR s rozpoznáním bloků textu

Software pro převod obrázku na textové objekty musí být schopen detekovat sektory na stránce, které text obsahují. Protože knižní obsahy mohou mít velice různorodou strukturu i layout, různé členění do kapitol a podkapitol, různé typografické i grafické provedení, může se stát, že automat určí tyto sektory chybně. Proto musí software umožnit manuální zásah a měl by být schopen se těmto novým situacím učit, aby pracoval stále dokonaleji.

Eliminace cizích bloků

Na stránce s obsahem je často přítomen i jiný text a různé grafické prvky, které je třeba rozeznat a odstranit ze zpracování. Pro další kroky je třeba již pracovat pouze s textovými bloky, ve kterých jsou obsaženy položky obsahu.

Rozlišení textových a číselných bloků, odstranění stopslov

Knižní obsah se skládá z položek obsahu, kde začátek tvoří název kapitoly a konec číslo stránky, na které daná kapitola v knize začíná. Číslo stránky následující kapitoly navíc určuje rozsah stran, který můžeme využít pro výpočet vah nebo zobrazení tematických mraků. Software proto musí z každé položky obsahu získat údaje o názvu položky a o umístění (stránce). I zde může situaci značně komplikovat různorodý layout obsahů.

U položek obsahu jsou často použita slova, která označují oddíly a kapitoly pouze formálně a nenesou tak žádné informace o obsahu jako takovém. Příkladem jsou označení jako „Kapitola 2: ", „Kniha první", „Rejstřík", či formální číslování jako „2.1.3". Tato formální označení je třeba rozeznat a z výsledku odstranit.

Textová analýza s důrazem na kontext položek obsahu

Tento krok je z celého postupu nejnáročnější a klade vysoké nároky na funkcionalitu softwaru. Odborné monografie používají ve většině případů víceúrovňové kapitoly, které jsou do sebe vnořovány a dědí z nadřazených kapitol kontext. Např. název podkapitoly „Cesty a jejich popis" lze interpretovat různě. Jiný význam bude mít v turistických cestopisech a jiný např. v pro-

gramování. Když však dáme tento název do kontextu s nadřazenou kapitolou „Práce se souborovým systémem", dokážeme již význam slova „cesta" interpretovat přesněji: jedná se o cesty v souborovém systému. Když to rozšíříme ještě o další kontext, vyplývající z názvu celého díla „Programujeme v jazyce Python", můžeme pak význam slova „cesty" určit zcela přesně.

Software proto musí umět pracovat s položkami obsahu nejen v přesném pořadí, ale musí být schopen rozlišovat nadřazenost a podřazenost položek, i míru jejich vnoření. Automatická detekce těchto vztahů obvykle vychází z layoutu položek na stránek (vnořené položky jsou často odsazeny) nebo typografického provedení (vnořené položky jsou často provedeny menším nebo méně výrazným písmem než položky nadřazené). Podřazené položky je třeba rozšířit o údaje z nadřazených položek a vytvořit tak plnohodnotná a jednoznačná klíčová slova.

Získání klíčových slov

Klíčová slova získaná z položek obsahu je možné ještě podle požadavků zformalizovat, vyloučit duplicity apod. Pokud zpracováváme monografie z více světových jazyků, můžeme v této fázi také volat automatické překlady apod. Lze také smazat údaje o kapitolách, které jsou dle nastavení marginální (např. když se dané kapitola rozkládá na méně než 2 stránkách, lze předpokládat, že je takové téma uváděno pouze okrajově a nedostatečně). Možností zpracování vydolovaných klíčových slov je mnoho.

Využití

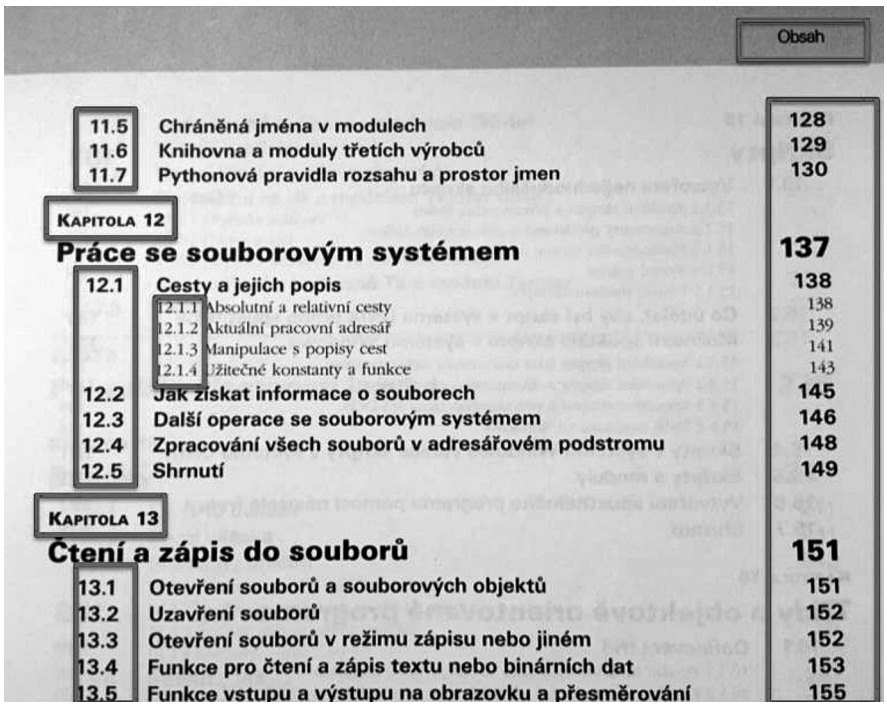
Výsledkem celé metody je seznam klíčových slov, které jsou připraveny na výstupu k využití informačními systémy, typicky knihovním systémem nebo indexem discovery systému. Seznam lze uložit přímo do vybraného pole bibliografického záznamu, který je následně indexován pro vyhledávání nebo sklízen pro indexování v jiném systému. Zajímavou možností je ukládání seznamů klíčových slov do sdíleného úložiště, který mohou využívat další knihovny, podobně jako se tak děje s obálkami knih v projektu obalkyknih.cz. Klíčová slova pak může lokální systém stáhnout ze sdíleného úložiště pomocí jednoznačného identifikátoru, např. ISBN.

Ve spojení s bibliografickými záznamy nám mohou klíčová slova z obsahu pomoci detekovat trendy v tématech, o kterých se za určité období nejvíce píše. Toho lze využít pro řadu zajímavých náhledů. Např. zobrazovat časovou osu, na které uživatel jasně vidí, jak často se téma vyskytuje v literatuře. Může také zobrazit, jak se dané klíčové slovo v čase vyskytuje v různých oborech, z čehož lze vystopovat prolínání témat mezi obory apod. Dalším využitím je zobrazování žebříčků nejpopulárnějších témat, což se může stát zajímavou alternativou ke klasicky pojímaným seznamům knižních novinek.

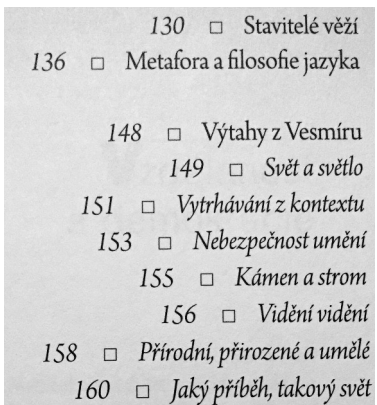
O projektu

Vývojem nástroje pro automatické rozpoznávání a indexování knižních obsahů se zabývá Národní technická knihovna v Praze. V letošním roce, v rámci projektu VISK, bude vyvinut prototyp nástroje, na kterém bude možné testovat zpracování obsahů monografií v českém jazyce. V případě zájmu o spolupráci v této oblasti nás můžete kontaktovat.

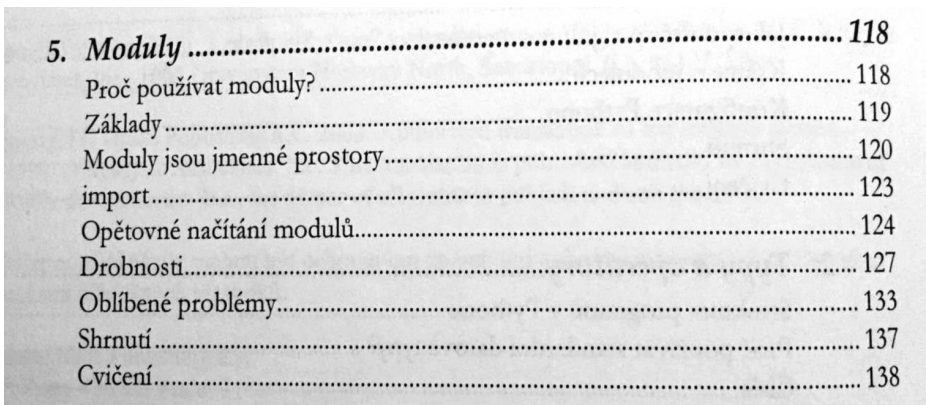
Příklady



Příklady stopslov, která netvoří klíčová slova



Příklad atypického obsahu, kdy je číslo stránky na začátku



Příklad podkapitol, jejichž význam určuje až kontext kapitoly a názvu díla

PhDr. Jan Pokorný, Ph.D.
jan.pokorny@techlib.cz
(Národní technická knihovna)