VYHLEDÁVÁNÍ V DATABÁZÍCH PLNÝCH TEXTŮ

Mgr. Vlastimil Červený
ÚISK FF UK
E-mail: vlastimil.cerveny@ff.cuni.cz

Úvod

Databáze plných textů jsou stále častějším a využívanějším zdrojem informací. Možnosti jejich budování a využívání, které jsou závislé na rozvoji výpočetní techniky a jejích periferií, jsou dnes značně rozsáhlé. Plnotextové databáze a obecně databáze primárních zdrojů (někdy též nazývané faktografické), které obsahují nejen texty dokumentů, ale také grafické, zvukové i obrazové záznamy, jsou dnes díky enormnímu poklesu cen paměťových médií a zároveň zvýšení jejich kapacity dostupné prakticky na všech úrovních informační práce. Prosazuje se rovněž trend používání plnotextových databází i v on-line dialogových službách na místech, kde bylo dříve možno použít pouze databáze bibliografické povahy. Využívání databází plných textů v on-line režimu bylo umožněno nejen zvýšením kapacity paměťových médií, ale i zvětšením propustnosti komunikačních sítí, které pomocí nejrůznějších technologií (např. FDDI) zaznamenávají obrovský nárůst přenesených dat. Klíčovou roli zde samozřejmě hraje Internet, jehož rozvoj je hnací silou vývoje a využívání těchto databází.

Na tomto místě je třeba poznamenat, že snahy o vytváření databází plných textů dokumentů jsou patrné i v samých počátcích dialogových informačních služeb. Jako příklad je možno uvést databázi plného znění deníku New York Times, která vznikla již v roce 1969[9]. Nejčastější uplatnění databází plných textů nalezneme právě v oblasti textů rozličných novin a časopisů, přičemž již dnes je možno na síti Internet zaznamenat aktivity směřující k vytváření elektronických knihoven obsahujících nejen bibliografické údaje, které je možno prohledávat pomocí on-line katalogů podobně, jako je tomu v klasické knihovně, ale i plné texty vyhledaných knih. Velmi časté uplatnění plnotextových systémů je rovněž v oblasti automatizovaných systémů právních dokumentů[4].

Dalším fenoménem úzce spjatým s databázemi plných textů a sítí Internet je *elektronické publikování,* které již dnes často dubluje nebo dokonce i nahrazuje klasické publikační techniky. Jako příklad je možno uvést nepřeberné množství elektronických mutací novin a časopisů dostupných na Internetu nebo dokonce periodik, která jsou publikována pouze elektronicky. Teoretickou možností je pak vytvoření globální elektronické knihovny všech knih v elektronické podobě. Síť Internet jako celek je možno považovat za jedinou distribuovanou databázi plných textů.

Nezanedbatelnou roli ve vývoji plnotextových systémů sehrála rovněž ekonomická stránka. Je třeba si uvědomit, že konkurentem plnotextových systémů jsou systémy, které obsahují indexované dokumenty s abstrakty a klíčovými slovy. Vlastní indexace dokumentů je intelektuálně a finančně velice náročná. V době, kdy bylo nutno dokumenty do plnotextového systému zpětně převádět z tištěné formy do počítačem čitelné podoby, byl tento proces oproti klasickému indexování značně ekonomicky neefektivní. K radikální změně a obrácení tohoto poměru došlo až v době, kdy byla většina tištěných dokumentů k dispozici i v elektronické podobě, která vznikla při jejich tvorbě. Dnes jsou prakticky všechny tištěné materiály připravovány pro tisk prostřednictvím výpočetní techniky, a proto jsou dostupné i elektronické verze tištěných dokumentů.

Trend používání elektronických dokumentů v informačním procesu je rovněž podpořen vznikem softwarových řešení pro tuto oblast, zejména pak vznikem speciálních formátů, které dovolují dokumenty komunikovat nezávisle na platformě. Takovým formátem, který je bezesporu nejpoužívanější v této oblasti, je formát PDF (Portable Document Format). Tento formát zachovává dokument v takové grafické kvalitě, ve které byl vytvořen. Jeho možnosti jsou již dnes využívány některými producenty různých tiskovin. Uživatel pak získává elektronický dokument, který je věrným obrazem dokumentu vytištěného na papíře (s přesným rozložením jednotlivých komponent dokumentu, textu, grafiky atd.). Další velmi výhodnou vlastností dokumentů ve formátu PDF je možnost spolupráce s různými systémy pro DTP (Desk Top Publishing). Vzhledem k tomu, že valná většina všech tištěných informací vzniká technologií DTP, ze které lze primární texty snadno přenést do některého plnotextového databázového systému, bude dnešní trend elektronického publikování ať už na CD-ROM, nebo na Internetu dále gradovat. Dokumenty ve formátu PDF jsou rovněž zaindexovatelné do běžných systémů pro tvorbu plnotextových databází. Takto vybudovaný databázový systém odstraňuje zásadní nevýhodu čistě textové databáze, neboť zpřístupňuje dokumenty, které mohou obsahovat i grafické a jiné komponenty, a vytváří tak faktografický databázový systém. Je třeba si uvědomit, že mnoho dokumentů obsahuje jak textové informace, tak informace ve formě grafů, obrázků a tabulek. V případě budování čistě textové databáze se texty získané z dokumentů, které obsahují i grafiku, stávají neúplnou faktografickou bází a uživatel je nucen podle bibliografických údajů (jsou-li k dispozici) vyhledat primární dokument.

Elektronické publikování a procesy s ním těsně spjaté ovšem nejsou jediným hnacím motorem tohoto odvětví, další hybnou silou je zefektivňování práce s dokumenty ve firemní praxi. Využívání informací jako strategické suroviny se stalo nezbytnou součástí podnikání v moderním globalizovaném světě. Vnitrofiremní a mezifiremní komunikace, využívání externích informačních zdrojů, průzkum trhu, to vše jsou činnosti, jež jsou

závislé na zpracování dokumentů. Rovněž je třeba vzít v úvahu, že většina dokumentů ve firemní a obchodní praxi není nijak významně a hlavně jednotně strukturována pro použití v nějakém relačně orientovaném databázovém systému. Jediným řešením pro možnost využití informací obsažených v dokumentech, které mají společné alespoň to, že je lze převést na textový formát, je jejich zařazení do plnotextové databáze.

V posledních několika letech je možno hovořit o vytvoření nového průmyslu, zabývajícího se dokumentovými informačními systémy. Tento segment informačního průmyslu má značný potenciál a roste velice dynamicky. D. C. Blair[1)] uvádí studii společnosti Delphi Consulting Group, Inc., která hovoří o 35% růstu v letech 1992-1995.

Současně s rozvojem technických prostředků pro tvorbu a využívání databází plných textů vzniká zájem o metody a techniky *vyhledávání* v těchto databázích. Následující převzatá tabulka[8)] ukazuje srovnání počtů článků vyskytujících se v databázích ERIC, INSPEC, Compendex Plus, LISA a Information Science Abstracts týkajících se vyhledávání v databázích plných textů. Přičemž je nutno poznamenat, že většina z těchto dokumentů má spíše deskriptivní charakter a jen malé procento z nich jsou původní výzkumné práce. Mezi často citované práce zabývající se výhradně vyhledáváním dokumentů v databázi plných textů patří články autorů Blaira a Marona[2, 3)].

Tab.: **Články o vyhledávání v databázích plných textů 1976-1995**

Databáze	76-80	81-85	86-90	91-95	Celkem
Eric	8	29	43	61	178
INSPEC	10	69	205	197	486
Com. Plus	2	5	9	26	42
LISA	49	102	306	117	578
ISA	1	17	57	33	108

Z údajů v tabulce je možno vysledovat vzrůstající trend i přes relativní pokles počtu článků o tomto tématu v období let 1991-1995. Vzrůstající zájem o tuto problematiku dokumentuje rovněž každoroční pořádání zvláštní konference TREC (Text REtrieval Conference) a vlastního výzkumu pod záštitou NIST (National Institute of Standards and Technology).

Generace plnotextových systémů

Formulace dotazu do databáze plných textů závisí na konkrétním softwarovém prostředku, který zpřístupňuje vlastní texty. Tyto prostředky se vyvíjely v silné závislosti na rozvoji informačních technologií, zvláště pak hardwaru. Systém vyhledávání hraje u plnotextové databáze analogickou úlohu jako selekční jazyk v knihovnickém systému, s tím rozdílem, že není určen k zaznamenání obsahu dokumentu, ale pouze k vyjádření selekčního požadavku uživatele.

Plnotextové systémy je dnes možno rozdělit na tři základní druhy podle způsobu vyhledávání[10)].

Systémy 1. generace

Systémy první generace determinované nedostatečným výpočetním výkonem tehdejších počítačů (první experimentální plnotextové systémy vznikaly již v padesátých letech[1)]), je možno charakterizovat jednoduchým vyhledáváním slov a jejich primitivních derivací, které ovšem nevycházely z lingvistického aparátu, jenž by je byl schopen odvozovat gramaticky, nýbrž konstruovaly derivace slov čistě mechanicky, pomocí jednoduchého maskování, nejběžněji pomocí pravostranného rozšíření slov. Vyhledávání pomocí pravostranného rozšiřování slov je běžné i dnes vzhledem k jednoduchosti jeho použití a k nenáročnosti na systémovou výbavu. Efektivita takového jednoduchého vyhledávání je ovšem velice nízká, neboť mnoho slov, která mají stejný kořen, má zcela jiný význam, což je možno demonstrovat na následujícím příkladě: Chceme-li nalézt všechny dokumenty obsahující slovo banka nebo jeho gramatické odvozeniny jako např. bankovní, zadáme pravostranné rozšíření slova *bank** . Budou tak nalezeny dokumenty obsahující slova *banka, banky, bankovní, bankéř* atd., což se jeví být v pořádku. Kromě toho však budou vybrány dokumenty obsahující např. slovo *banket,* které s danou problematikou zjevně nijak nesouvisejí.

Systémy 1. generace neumožňují vyhledávat kombinace několika slov za použití dalších operátorů. Kromě slov nerozlišují další části textu, jako jsou věty, odstavce, stránky dokumentu apod. Se stoupajícím výkonem výpočetní techniky stoupaly nároky uživatelů na výkonnost samotného vyhledávacího systému. Hnací silou dalšího rozvoje byly projekty automatizace knihovnických a bibliografických systémů za pomoci výpočetní techniky. Takový automatizovaný systém musel mít schopnost zpracovávat podmínky pro selekci dokumentů, vyjádřené pomocí booleovských spojek, a použít je na vyhledání bibliografického záznamu. Booleovský model vyhledávání byl tedy vyvinut pro práci s bibliografickými databázemi, které mají určitá specifika oproti databázím strukturovaným nebo databázím plných textů. Hlavní specifikum těchto bází tkví v tom, že údaje v nich jsou převážně textové povahy a že jednotlivé položky bibliografické databáze mohou obsahovat více údajů (např. položku autor nebo položku klíčová slova). Obecně je tento problém řešitelný i v rámci dnes běžného relačního modelu dat. Ovšem vznik i praktické uplatnění tohoto systému je datován až po rozvoji booleovského modelu.

Systémy 2. generace a studie STAIRS

Systémy druhé generace je možno charakterizovat možností vyhledávání slov a slovních spojení pomocí

booleovských a proximitních operátorů. Použití booleovských spojek AND, OR a NOT a proximitních operátorů přináší možnost vyhledávat slova nebo slovní spojení v zadané vzdálenosti od sebe nebo v jedné větě či odstavci. Pro vyjádření hierarchických a prioritních vztahů je v těchto systémech možno použít závorky. U těchto systémů rovněž začínají vznikat možnosti jednoduchého vyhledávání pomocí automatických gramatických derivací. Ve srovnání se systémy 1. generace je zdokonalena rovněž možnost rozšiřování slov. Zůstalo pravostranné rozšiřování slov (tzv. sufix), které pracuje shodně jako u předchozích systémů, a přibyla možnost levostranného rozšíření slov (tzv. prefix), které pracuje analogicky pravostrannému rozšíření. Často je k dispozici rovněž možnost maskovat určitou pozici ve slově libovolným znakem.

Zásadní nevýhoda systémů druhé generace spočívá v použití dvouhodnotové logiky k vyhodnocování dotazů. Na základě booleovské algebry je dokument buď vybrán, nebo nevybrán - jiná možnost zde není. Díky této vlastnosti dvouhodnotového systému jsou přesnost a úplnost vyhledávání konfliktní vlastnosti. V čistě booleovském systému rovněž chybí jakákoli možnost automatického hodnocení relevance vyhledaných dokumentů. Systém není schopen seznam vyhledaných dokumentů seřadit podle nějakého kritéria, které by hodnotilo relevanci dokumentu. Existuje zde pouze možnost řazení podle časové řady apod.

Připomeňme definici koeficientů přesnosti (relevance) a úplnosti, které slouží k měření efektivity vyhledávání v dokumentografických systémech. Koeficient přesnosti P (Precision) je určen jako poměr počtu vybraných relevantních dokumentů k počtu všech vybraných dokumentů. Koeficient přesnosti P tedy určuje, jak dobře systém vyhledá jen relevantní dokumenty.

Koeficient úplnosti R (Recall) je určen jako poměr počtu vybraných relevantních dokumentů ku počtu všech relevantních dokumentů. Koeficient úplnosti R měří, jak dobře systém vyhledá všechny relevantní dokumenty.

Oba koeficienty jsou ve vztahu nepřímé úměrnosti, což lze nahlédnout rovněž z jednoho z principů tradičního učení o pojmu a jeho rozsahu; platí totiž princip obráceného poměru rozsahu a obsahu[7)].

Koeficienty úplnosti a relevance ovšem není možno považovat za nějakou absolutní míru, neboť samu relevanci dokumentu většinou definujeme jako míru užitečnosti, kterou uživatel připisuje získanému dokumentu při řešení určitého problému. Je zřejmé, že záleží na hodnotiteli nebo na samotném uživateli, jaká je vlastně míra relevance daného dokumentu. Je to tedy do jisté míry subjektivní záležitost a velmi záleží na zvolené metodologii hodnocení dokumentů jako relevantních.

Určitá neobjektivita při procesu měření účinnosti vyhledávacího systému je bohužel neodstranitelná, což je dáno tím, že relevantní informace v systému uložené jsou „zakódovány" v textu a je tedy velmi složité je nějakým způsobem přesněji lokalizovat. Podílí se na tom rovněž složitost jazyka a tím i světa, ve kterém se v tomto případě pohybujeme.

Použití booleovské algebry a proximitních operátorů znamenalo mohutný skok v technologii vyhledávání dokumentů. Poskytovatelé on-line informačních a vyhledávacích služeb tak dostali do ruky poměrně silný nástroj pro vyhledávání informací ve strukturovaných dokumentech. Dnes pracují systémy druhé generace prakticky ve všech komerčních databázových systémech typu DIALOG a DATA-STAR, často bývají rozšířeny o tezaurovou podporu. Booleovská algebra v kombinaci s tezaurem, který definuje vztah nadřazenosti, podřazenosti a asociativní vztahy, je poměrně silný nástroj pro vyhledávání ve strukturovaných databázích. Nejčastější využití mají tyto systémy při vyhledávání v databázích sekundárních dokumentů (strukturovaných).

Trend rozvoje databázových systémů jde ovšem směrem od databází strukturovaných, obsahujících sekundární dokumenty, k databázím faktografickým, obsahujícím plné texty vlastních dokumentů. Poměrně dlouhou dobu byly i pro vyhledávání v databázích plných textů užívány systémy druhé generace. Čistý booleovský model bývá v těchto systémech rozšířen o některé funkce a operátory, které mají zvýšit efektivitu vyhledávání dokumentů a rovněž nabízejí seřazení dokumentů podle skóre relevance, stanoveného systémem. Tyto systémy ovšem nestanovují skóre relevance na základě vnitřních gramatických a obsahových analýz, nýbrž se snaží odhadnout relevanci v závislosti na počtu vyhledávaných slov v dokumentu, např. vzhledem k jeho délce a podobně. Objevují se zde proximitní operátory typu NEAR, PHRASE, PARAGRAPH, které vyhledávají požadovaná slova nebo slovní spojení v určité vzdálenosti od sebe v určité části textu (věta, odstavec). Některé systémy druhé generace, které bychom pro tuto vlastnost nazvali systémy generace dva a půlté, umožňují dokonce jednotlivým částem booleovského dotazu přiřadit váhy, podle kterých je pak vypočteno skóre relevance. Tímto způsobem je možno vyjádřit prostý, ale velmi důležitý fakt, že některá slova charakterizují dokument více a některá méně. Existují matematické metody, které dokáží pracovat s booleovskými spojkami, jimž jsou přiřazeny váhy důležitosti a které dokáží vyčíslit skóre relevance (vektorový model dokumentu). I přes tato rozšíření booleovského modelu nebyly výsledky vyhledávání ve velkých textových databázích uspokojivé.

Hybným impulsem nutnosti změny v přístupu k vyhledávání informací v dokumentech plných textů byla až studie STAIRS[2)] provedená roku 1985 na systému IBM/STAIRS. Byl to první experiment, který hodnotil koeficienty úplnosti a relevance systému STAIRS (Storage And Information Retrieval System), který obsahoval velké množství dokumentů. Z výše uvedených definic koeficientů úplnosti a relevance je patrné, že taková studie provedená nad velkou bází textových informací je velice finančně i systémově nákladná, neboť např. pro určení koeficientu úplnosti R je nutno zjistit počet dokumentů, které jsou relevantní, ale které systém za relevantní neoznačil, a proto nevyhledal. Počet těchto relevantních, ale nevyhledaných dokumentů není možno zjistit, aniž by byl znám obsah celého zkoumaného fondu.

Studie STAIRS měla za úkol zmapovat, respektive zhodnotit efektivitu vyhledávání informací v daném systému. Vyhledávání se týkalo databáze právnických textů. Měla za úkol zjistit standardní koeficienty úplnosti a relevance. Základem studie bylo zpracování a zhodnocení 50 dotazů provedených v databázi textů o rozsahu ekvivalentnímu 350 000 stran. Náklady na tuto studii byly vyčísleny na přibližně 500 000 $. Výsledky dotazů byly poměrně překvapující, neboť vykazovaly přesnost 80 %, ale úplnost pouhých 20 %[2]. Tento výsledek byl impulsem pro vývoj nového vyhledávacího systému třetí generace.

Zastavme se nejdříve u důvodů, které vedly k tak nízkému koeficientu úplnosti. Obecně si lze každý databázový systém představit jako model nějakého světa, který obsahuje nějaké objekty. Ke každému takovému modelu světa je možno definovat soubor propozic, kterým je množina tvrzení o momentálním stavu světa objektů. Tyto propozice jsou vyjádřitelné selekčním jazykem. Např. relační databáze má v nejjednodušším případě jazyk, který popisuje např. dvě položky - jméno pracovníka a výšku jeho platu. Selekční jazyk je v tomto nejjednodušším případě schopen zjistit momentální stav všech pracovníků vzhledem k jejich platu. Výrazy selekčního jazyka v relačních systémech vždy korespondují se strukturou a s vlastními uloženými daty. V plnotextové databázi je popisovaný svět příliš široký a mezi jednotlivými propozicemi existují velmi složité, často nepostřehnutelné vztahy. Rovněž jazyk, ve kterém jsou informace uloženy, je příliš složitý a nejednoznačný (synonymie, homonymie a polysémie). Dalším problémem je, že mnoho dokumentů je relevantních, aniž by to bylo nějak významově zřejmé. Mnohdy jsou dokumenty k danému tématu relevantní jen svou vlastní existencí, dejme tomu v nějaké významné časové posloupnosti. Např. pokud se v určitý čas objevil dokument s nějakým politickým prohlášením, je možné předpokládat, že tento dokument může být velmi relevantní pro mnoho skupin vyhledávaných témat. Důležitá je zde i časová posloupnost jednotlivých dokumentů. Další faktor, který se značně podílí na určování relevance dokumentů, souvisí s kognitivními procesy člověka v průběhu vyhledávání a hodnocení dokumentů. Víme, že pokud přijímáme informaci o nějakém fenoménu, pak jejím přijetím dochází i ke změně námi vnímaného a poznávaného jevu. Záleží tedy do značné míry na pořadí, v jakém uživatel dokumenty hodnotí, jsou-li pro něj relevantní či nikoliv. Z toho vyplývá závěr, že vyhledávání dokumentů v databázi plných textů byl mělo mít charakter procesu se zpětnou vazbou.

Jedním ze závěrů, který učinili Blair a Maron[2] je, že uživatel při formulaci dotazu nemůže znát všechny pojmy, jež jsou určující pro relevanci hledaného dokumentu. Objevuje se zde problematika definice pojmů, která je velkým problémem systematických selekčních jazyků v knihovnických systémech.

Systémy 3. generace

Systémy třetí generace je možno charakterizovat zcela novým přístupem k vyhledávání dokumentů, který je založen na principech:

1. rozkladu pojmu na podpojmy
2. vážení jednotlivých podpojmů (větví pojmového stromu)
3. neostrého vyhodnocování dotazů

Dotaz v systému 3. generace reprezentuje pojem, respektive ideu vyhledávaného tématu. Jádrem dotazu je stromová hierarchická struktura, která rozkládá hledané téma na podtémata a přiřazuje jednotlivým částem váhy, které vyjadřují do jaké míry příslušné podtéma přispívá k celkovému určení tématu. Systém je pak schopen vypočítat míru relevance (nejčastěji udávanou v % nebo hodnotou v intervalu 0,1), podle které řadí vyhledané dokumenty. Takovéto uspořádání má oproti předchozím systémům značné výhody, např. vyhledané dokumenty nejsou systémem hodnoceny podle dvouhodnotové logiky zda obsahují daný termín nebo nikoliv, je tedy možno vyhledávat neostře. Tato vlastnost s sebou přináší samozřejmě nutnost definice nových logických operátorů, které mají tyto neostré selekční charakteristiky, např. operátor ACCRUE v případě systému TOPIC. Další výhodou je rozklad pojmů na podpojmy v podobě hierarchického stromu, což značně zvyšuje přehlednost a umožňuje tvorbu velmi rozsáhlých dotazů. Stromová struktura dotazů umožňuje použití jednotlivých definovaných větví v jiných částech dotazu, čímž vzniká možnost použití parciální rekurze při konstrukci dotazu. Vážení pojmů a podpojmů a jejich uspořádání přináší do vyhodnocování dotazů využití „neostré“ fuzzy logiky. Při zpracování dotazů je pro každý dokument vypočítáno jeho celkové skóre relevance, které vyjadřuje, do jaké míry odpovídá zadanému dotazu.

Jedním z nejdokonalejších systémů pro vyhledávání plnotextových dokumentů je systém TOPIC americké firmy Verity, Inc. Systém TOPIC je jedním z pěti existujících komerčních systémů, které jsou označovány za pojmově orientované vyhledávací systémy („concept based retrieval“) podle již zmiňované analýzy firmy Delphi Consulting Group, Inc.[1].

Tento pojmově orientovaný vyhledávací systém se opírá o možnost definice pojmu pomocí hierarchické stromové struktury. Využívá se zde východisek tradičního učení o pojmu a jeho rozsahu a celý proces je do značné míry analogický tvorbě hierarchického selekčního jazyka. Definovaný pojem (v tomto případě hovoříme o topiku) je reprezentován názvem celé stromové struktury, který tvoří její kořen. Jednotlivé větve stromové struktury představují podpojmy, které jsou analogické jednotlivým podtřídám u hierarchického selekčního jazyka typu MDT. Jednotlivé větve vytvářeného topiku jsou dále rozložitelné na další podvětve analogické dalším podtřídám nadřazených tříd. Při konstrukci topiku postupujeme podle logických pravidel, která jsou běžná i při

klasifikaci do tříd v knihovnické praxi. Je nutno tedy splnit požadavek na disjunkci podtříd stejné úrovně, dodržovat hierarchickou strukturu tříd apod.

Systém TOPIC eliminuje jeden z nedostatků booleovských vyhledávacích systémů, kterým je přílišná ostrost operátoru AND, jenž nevyhledá dokument, pokud neobsahuje všechna slova tímto operátorem spojená, zavedením operátoru ACCRUE. Na následujícím příkladě si ukážeme, jak tento nový operátor pracuje. Z naměřených charakteristik je možno získat principiální představu o jeho funkci, nicméně přesný matematický popis výpočtu koeficientu relevance tohoto operátoru není v běžné literatuře dostupný a je zřejmě předmětem obchodního tajemství. Obrázek č. 1 ukazuje definici jednoduchého topiku, který spojuje dva pojmy stejné úrovně (oba reprezentují názvy dvou měst) pomocí operátoru ACCRUE, přičemž slovu Praha je přiřazena váha 0.6 a slovu Brno váha 0.4. Operátor ACCRUE pracuje zjednodušeně řečeno tak, že se nejdříve chová jako operátor AND a po nalezení všech dokumentů vyhovujících této podmínce se začne chovat jako operátor OR. Dotaz byl učiněn v databázi plného textu časopisu Ekonom, ročník 1998, ve firmě Economia, a.s. Tato databáze obsahuje 5420 článků.

Obr. 1

Následující grafy zobrazují skóre relevance a počty vyhledaných dokumentů tak, jak je vyhledá operátor ACCRUE. Na grafu č. 1 můžeme vidět tři skokem ohraničené části.

Graf č. 1

První části systém přiřadil hodnotu 0.75 a představuje tu část operátoru ACCRUE, která se chová jako operátor AND. Další části grafu znázorňují situaci, kdy se začne chovat jako operátor OR, přesněji řečeno jako operátor logické funkce XOR. Zde je možno pozorovat, jakým způsobem se projeví definice jednotlivých vah v jednotlivých větvích topiku. Čára mající hodnotu relevance 0.6 reprezentuje Prahu, zatímco čára reprezentující Brno má hodnotu 0.4. Je tedy patrné, že operátor ACCRUE splňuje jak podmínku vysoké relevance, tak podmínku vysoké úplnosti vyhledávaných dokumentů. Tento operátor vyhledá stejný počet dokumentů jako dotaz založený na booleovském operátoru OR a zároveň na první místa seznamu vyhledaných dokumentů umístí dokumenty relevantní k dotazu s booleovským operátorem AND. Dotaz zkonstruovaný tak, jak je zobrazeno na grafu č. 1, nám dává seznam dokumentů, které jsou rozděleny do tří skupin relevance 0.75, 0.6, 0.4. Vyhledaných dokumentů je však více než tisíc a bylo by tedy potřeba tento seznam seřadit jemnějším způsobem. U čistě booleovských systémů však není možnost systémově stanovit koeficient relevance, neboť ten je vždy roven jedné. V části týkající se rozšíření booleovského modelu jsme některé možnosti řešení tohoto problému naznačili. V systému TOPIC k tomuto účelu slouží modifikátor MANY, který přiřadí relevantnímu dokumentu hodnotu z intervalu <0,1> podle hustoty výskytu hledaného slova nebo fráze. Operátor MANY určuje hustotu výskytu slova v dokumentu, nikoliv prostý počet výskytů. Hustota je definována jako počet výskytů v závislosti na délce textu, může se tedy stát, že dlouhý dokument, který obsahuje více výskytů hledaného řetězce, může mít menší skóre relevance než kratší text, který obsahuje méně výskytů. Graf č. 2 ukazuje výsledky stejného dotazu jako graf č. 1 s použitím modifikátoru MANY.

Graf č. 2

Graf č. 3 ukazuje pro srovnání chování systému TOPIC při nejjednodušším dotazu s využitím operátoru ACCRUE s defaultně nastaveným modifikátorem MANY a bez definice vah jednotlivých větví.

Graf č. 3

Průběh tohoto grafu se do značné míry blíží křivce, která vyjadřuje obecný vztah mezi koeficienty úplnosti a relevance. Skok mezi jednotlivými částmi grafu ohraničuje chování typu AND a OR.

Kromě použití operátoru ACCRUE je pro systém TOPIC určující způsob budování jednotlivých topiků. Výše uvedený příklad byl triviální, protože měl za cíl objasnit funkci operátoru ACCRUE.

Nyní si ukažme nějaký složitější topik. Je vidět z následujícího obrázku č. 2, který zobrazuje pro představu definici prázdného topiku instituce[6].

Obr. č. 2

Pomocí stromové struktury je možno vytvářet značně složité definice pojmů. V zásadě existují dvě základní strategie tvorby: buď od obecného k jednotlivému (Top-Down Design), nebo od jednotlivého k obecnému (Bottom-Up Design). Jak je vidět z předchozího obrázku, topiky mohou mít poměrně komplikovanou strukturu, proto je výhodnější při jejich konstrukci postupovat druhou naznačenou metodou, neboť nastavit váhy u jednotlivých podvětví již nadefinovaného topiku a uchovat si zároveň představu o vlivu jednotlivých vah na výsledek vyhledávání je prakticky nemožné. Zároveň jsme výše určili jako jednu z podmínek kvalitního vyhledávání dokumentu možnost použití zpětné vazby při konstrukci dotazu. Proto je vhodné topiky konstruovat po částech, u kterých je potřeba v několika iteracích doladit hladiny vah jednotlivých větví.

Vzhledem k tomu, že lze na topik odkazovat jeho jménem nebo kombinovat jeho jednotlivé větve a jednotlivé topiky slučovat pod ještě obecnější topiky, rýsuje se zde možnost vytvoření selekčního jazyka založeného např. na bázi MDT. Jednotlivé topiky by v takto vybudo-

vaném informačním systému reprezentovaly výrazy selekčního jazyka. Celý systém by představoval automatizovaný systém vyhledávání dokumentů pomocí předem definovaných topiků. Je třeba si uvědomit, že selekční systém je v tomto případě zcela nezávislý na faktickém obsahu dokumentů uložených v databázi. K selekci a indexaci v knihovnickém smyslu by tak docházelo až v procesu vyhledávání dokumentů.

Již dnes existují funkční systémy na automatizované třídění přicházejících dokumentů, např. agenturního zpravodajství. Dokumenty přicházejí do systému, kde jsou automaticky podrobeny selekci pomocí dobře nadefinovaných topiků (politika, ekonomika apod.). Z výše uvedeného je patrné, že klíčovým aspektem úspěšnosti podobného plnotextového systému je vlastní vyvážená definice topiků. Je zřejmé, že je to práce pro specialistu, srovnatelná s tvor-bou expertních systémů, neboť dobře nadefinovaná báze topiků představuje vlastně bázi znalostí.

Další vývoj těchto systémů bude pravděpodobně založen na nejnovějších poznatcích moderní logiky, lingvistiky a umělé inteligence. V moderní logice jsou to zejména teorie, pokoušející se znovu definovat a pevně zakotvit vlastní pojem „pojmu“ (viz např.[7]). Z hlediska databáze plných textů je z moderní lingvistiky velice zajímavý směr, který se nazývá „textová lingvistika“ [5]. Jedná se o lingvistickou disciplínu, která považuje za základní jednotku jazyka text. Na vývoji moderní lingvistiky je zajímavé, jak se postupně přenáší zájem jazykovědců ke zkoumání stále větších celků, od hlásek, přes věty až k celým textům (další pravděpodobný krok bude zřejmě od textu k hypertextu). Textová lingvistika již definuje některé pojmy sloužící k popisu textu jako celku. Některé z nich (Makrostruktura, Témata) nápadně korespondují s definicí topiku v systému TOPIC. Dalším směrem ve vývoji těchto systémů je aplikace umělé inteligence, zejména pak systému na porozumění přirozenému jazyku. Informační systém, který by byl založen na tomto principu, by nepotřeboval selekční jazyk a vyhledávání dokumentů by probíhalo dotazováním se systému v přirozeném jazyce.

Použitá literatura:

1) BLAIR, D. C. STAIRS Redux: Thoughts on the STAIRS Evaluation, Ten Years after. *Journal of the American Society for Information Science*, 1996, Vol. 47, no. 1, s. 4-22.

2) Blair, D. C. , MARON, M. E. An evaluation of retrieval effectivness for a full-text document-retrieval system. *Communications of the ACM*, 1985, Vol. 28, no. 3, s. 289-298.

3) BLAIR, D. C. , MARON, M. E. Full-text information retrieval : further analysis and clarification. *Information Processing & Management*, 1990, Vol. 26, no. 3, s. 437-447.

4) KNAPP Viktor, CEJPEK Jiří. *Automatizované vyhľadávanie informácií v právnych textoch*. Bratislava : Slovenská technická knižnica, 1980. 169 s.

5) ČERNÝ, Jiří. *Úvod do studia jazyka*. Olomouc : Rubico, 1998. 248 s.

6) DÍTĚ, Jan. *Návrh báze dat pro vnější zadávání a údržbu témat systému TOPIC* : diplomová práce. Praha : VŠE, 1997. 74 s.

7) MATERNA, Pavel. *Svět pojmů a logika*. Praha : Filosofia, 1995. 131 s.

8) SIEVERT, M. C. Full-Text Information Retrieval : Introduction. *Journal of the American Society for Information Science*, 1996, Vol. 47, no. 4, s. 261-262.

9) VLASÁK, Rudolf. *Světové informační systémy a služby - Informační průmysl*. Praha : Karolinum, 1993. 178 s.

10) ŽBIRKA, Jan. Vyhledávání v úplných textech ekonomických periodik metodami 3. generace. In: *CS ONLINE 95 - zborník.* Bratislava, 1995, s. 71-74.

Pozn.: *Příspěvek je pracovním materiálem pro autorovu disertační práci.*