Úskalí vyhledávání pomocí příkazového jazyka (CCL) v databázi NKC Národní knihovny ČR

Otakar Pinkas
VŠE Praha

Úvod

Tento článek vznikl na základě ověřování výsledků semestrálních úkolů studentů informatiky na VŠE v Praze. Jejich úkolem byla rešerše ze zvoleného informačního systému přístupného pomocí protokolu http. Rešerše měla být přesná a úplná a neměla obsahovat jen klasické operátory and, or a not. Řada studentů si zvolila databázi NKC, protože oceňovala její obsah, rozsah a vyhledávací možnosti programového systému ALEPH. Někteří z nich narazili na problémy, o kterých článek pojednává, avšak raději změnili dotaz, než aby analyzovali příčiny neočekávaného chování systému. Nebo se o svých problémech sice zmínili, ale dále je neanalyzovali. Část z nich raději přešla na systém Proquest 5000 a úkol řešila v něm.

Hodnocení úkolů vyvolalo nutnost samostatného ověření výsledků rešerší. Testovali jsme proto různé typy dotazů v jazyce CCL v databázi NKC. Naší snahou bylo zjistit, které typy dotazů přinášejí spolehlivé výsledky a které nikoli. V případě selhání systému jsme se snažili určit obecné podmínky selhání. Jedná se samozřejmě o pohled uživatele, pohled programátora vyžaduje volný přístup k datům a programům. Časově byly testy rozloženy na jaro a podzim r. 2004 a začátek r. 2005.

Užívání jazyka CCL bylo určeno obecnou i specifickou nápovědou systému. Na základě zkušeností z testů a porovnáním s jinými systémy jsme poukázali na možnosti zpřesnění a zvýraznění nápovědy tak, aby více pomáhala uživateli. V zájmu objektivity musíme vedle selhání systému zavést pojem „selhání uživatele“. Selháním uživatele je míněn jeho nesprávný výklad nápovědy nebo neoprávněné užívání některých typů dotazů dotazovacího jazyka. Právě vyhledávání pomocí CCL v databázi NKC vyvolává nejistotu, zda jde o selhání uživatele nebo systému.

Příkazový jazyk CCL přináší oproti vyhledávacím formulářům bohatší možnosti vytváření přesnějších a úplnějších dotazů. Jeho výhody se projeví právě při vyhledávání v rozsáhlých databázích, tj. i v databázi NKC. Představuje určité úskalí, protože uživatel nemá k dispozici dostatečně přesný popis dotazovacího jazyka a protože příklady uváděné v nápovědě navozují představu snadného, rychlého a přesného vyhledávání.

Ověřování vyhledávacích možností jazyka CCL v databázi NKC

V nápovědě k systému ALEPH nenajdeme definici slova. **Slovem** budeme rozumět souvislý řetězec abecedních znaků. Seznam abecedních znaků může být vymezen např. tabulkou. Slovem může být např.: předložka „v“, číslo v textovém tvaru „1000“ nebo slovo „databáze“. Zvláštní pozornost věnujeme slovům českého jazyka, protože čeština se užívá k popisu dokumentu v různých údajích katalogizačního záznamu.

Jednoslovné dotazy

Uvádíme několik příkladů nejjednodušších dotazů a počty vyhledaných záznamů.

D1 Slova-Všechna pole= (lák) 264
D2 Slova-Všechna pole= (péro) 87
D3 Slova-Všechna pole= (bój) 3785
D4 Slova-Všechna pole= (kůl) 3200
D5 Slova-Všechna pole= (týl) 649

Analýzou vyhledaných záznamů a zpracováním modifikovaných dotazů zjistíme ekvivalenci některých hlásek pro vyhledávání, a to jak samohlásek, tak souhlásek. Liší se jen diakritickým znaménkem.

Krátké a dlouhé samohlásky

Krátké a dlouhé samohlásky v češtině: a, e, i, o, u, y - á, é, í, ó, ú, ů, ý. Písmeno „ě“ neoznačuje zvláštní samohlásku.

V databázích NKP se uvnitř systému nerozlišují slova, která se liší délkou samohlásky. Za **stejná** slova se **pro vyhledávání** považují např.: lak a lák, peče a péče, vir a vír, boje a bóje, uhel a úhel, luna a lůna, tyl a týl.

Hledáme-li záznamy odpovídající slovům „vir“ a „vír“, dostaneme vždy stejnou množinu výsledků. Obsahuje záznamy o virech a zároveň o vírech (vodních, vzdušných, atp.). V lístkovém katalogu budou množiny disjunktní.

D6 Slova-Všechna pole= (vir) 94 http://sigma.nkp.cz/F/... ?func=find-c&ccl_term=vir
D7 Slova-Všechna pole= (vir) 94 http://sigma.nkp.cz/F/... ?func=find-c&ccl_term=v%C3%ADr

Chceme-li jen záznamy obsahující slovo „vir“, musíme výsledný soubor **filtrovat**, a to ručně nebo s použitím filtrovací funkce. Textové editory rozlišují slova lišící se délkou samohlásky a umějí je vyhledat.

Krátké a dlouhé „i“ budou mít v URL dotazu různou reprezentaci (viz příklady výše), ale po vnitřní úpravě v systému ALEPH bude kódová reprezentace shodná. Dá se předpokládat **převod** „í“ z kódu UTF-8 (dva bajty) na jednobytové krátké „i“.

Souhlásky

B, c, č, d, ď, f, g, h, ch, j, k, l, m, n, ň, p, r, ř, s, š, t, ť, v, z, ž.
Dvojice: c, č - d, ď - n, ň - r, ř – s, š - t, ť – z, ž

V listopadu 2004 jsme testovali v databázi NKC, zda systém odlišuje pro vyhledávání měkké a tvrdé souhlásky či nikoli. Lze konstatovat, že v některých dvojicích ano a v jiných nikoli. V úvahu jsme vzali všechna slova, bez omezení na pole. Výsledky ukazuje tabulka.

Slovo – bez měkčení	Počet záznamů	Slovo – s měkčením	Počet záznamů	Datum vyhledávání
kocka	70	kočka	348	
Dák	5	ďák	5	15.2.2005
kunka	3	kuňka	3	
prečan	53	přečan	0	15.2.2005
troska	47	troška	24	15.2.2005
tápal	1	ťápal	1	
Zak	265	žák	840	15.2.2005

V dotazu „kuňka“ dostaneme kromě záznamů o žábách také záznam, jehož autorem je Kunka. Slova „kuňka“ a „kunka“ vedou ke stejné množině vyhledaných záznamů, tj. písmena „n“ a „ň“ jsou pro vyhledávání ekvivalentní.

Lze konstatovat, že dvojice souhlásek c, č – r, ř – s, š – z, ž se pro vyhledávání berou jako odlišná písmena, ostatní jako shodná.

Část slova s rozšířením zprava

Část slova s rozšířením zprava – slovo zkrácené zprava o žádné, jedno nebo více znaků a doplněné zprava znakem pro rozšíření. Příklady: vir*, počítač*, databáz*.

Při vyhledávání může hrát roli stálé zkracování slova s pravostranným rozšířením ve vztahu k omezení v počtu vyhledatelných záznamů. **Přetečením** se myslí překročení stanovené horní hranice vyhledávání záznamů (práh vyhledávání). Informace o prahu vyhledávání v nápovědě nemusí odpovídat skutečnému stavu.

V dubnu 2004 vedlo hledání v databázi NKC k těmto výsledkům:

Mi*	přetečení	Micros*	1084
Mic*	přetečení	Microso*	871
Micr*	1342	Microsof*	863
Micro*	1337	Microsoft	863

D8 Slova-Všechna pole= (mic*) 0 http://sigma.nkp.cz/F/... ?func=find-c&ccl_term=mic*

Přetečení může, ale nemusí, sloužit k zjištění příčiny neočekávaného chování systému v případě dotazů s dvěma termíny. V některých případech totiž stačí, aby jeden operand dotazu vedl k přetečení, jindy nikoli. Stačí porovnat dotazy D9 a D39.

Systém reaguje na přetečení poněkud nekorektně. V seznamu vyhledaných záznamů se objeví výstražný trojúhelník se slovy „příliš mnoho záznamů“ a v historii dotazů se nastaví počet vyhledaných záznamů na 0. Právě indikace ve formě 0 je matoucí. V r. 2005 jsme se setkali i s případem, že se dotaz s přetečením nedostal do historie dotazování.

Dotazy o dvou termínech a jedné základní logické spojce

Budeme se zabývat jednoduchou formou dotazu, kterou lze vyjádřit tímto schématem:

[POLE=](R[*] AND/OR/NOT S[*])

Význam symbolů

POLE zastupuje některé z přípustných **logických polí** databáze: WRD, WTL, WAU, WPB,WPP, atd.
R, S označují slova nebo jejich části zkrácené zprava, hvězdička značí pravostranné rozšíření a hranaté závorky znamenají, že R, S mohou být slova s rozšířením nebo bez něho.
AND je logická spojka „a“, OR je logická spojka „nebo“ a NOT znamená spojku „a ne“.

Dotazy podle tohoto schématu nevyvolají, až na drobné výjimky, žádnou neočekávanou reakci systému. Dotaz D9 je právě tou výjimkou.

D9 Slova-Všechna pole= (mic* and c++) 0

Dotazy o dvou termínech a jednom operátoru vzdálenosti

Formu tohoto typu dotazu lze vyjádřit obdobně jako v předchozím příkladě, avšak místo základních logických spojek jsou použity vzdálenostní operátory. Schéma dotazu:

[POLE=](R[*] !n/%n S[*])

Význam symbolů je shodný s významem symbolů dříve uvedených. Nové jsou operátory vzdálenosti:
!n – termín R předchází před termínem S a v případě n=0 neleží mezi nimi žádný termín. Je-li n=1, pak mezi R a S může ležet nejvýše jeden termín, atd.
%n – termíny R a S mohou být v libovolném pořadí; je-li n=0, neleží mezi nimi žádný termín. Je-li n=1, pak mezi R a S (nebo S a R) může ležet nejvýše jeden termín, atd.

V nápovědě k databázi NKC najdeme speciální definici blízkosti. Slova R a S jsou **blízká**, jestliže R bezprostředně předchází S.
Slovní spojení „relační databáze“ je příkladem blízkosti dvou slov. V systému není zaveden speciální operátor pro takto definovanou blízkost. Ve vyhledávacím formuláři lze nastavit hodnotu vstupního pole „adjacent“ na „ano“ nebo „ne“.

Dotazy s operátorem vzdálenosti a bez rozšíření slov

Zdá se, že oba operátory vzdálenosti fungují správně pro slova bez rozšíření. Podle očekávání vede menší vzdálenost mezi slovy k menšímu počtu vyhledaných záznamů a naopak.

Závazné pořadí slov, slova bez rozšíření nebo maskování
D10 Slova-Všechna pole= (český !1 venkov) 10 D11 Slova-Všechna pole= (český !0 venkov) 9

Nezávazné pořadí slov, slova bez rozšíření nebo maskování

D12 Slova-Všechna pole= (český %5 venkov) 15 D14 Slova-Všechna pole= (český %3 venkov) 15
D13 Slova-Všechna pole= (český %4 venkov) 15

Sada obdobných dotazů z informatiky opět splňuje očekávání. Operátor závazné vzdálenosti funguje.
D15 Slova-Všechna pole= (tipů !4 c++) 1 D18 Slova-Všechna pole= (tipů !1 c++) 0
D16 Slova-Všechna pole= (tipů !3 c++) 0 D19 Slova-Všechna pole= (tipů !0 c++) 0
D17 Slova-Všechna pole= (tipů !2 c++) 0

V názvu vyhledané knihy předchází slovo „tipů“ před slovem „C++“ a mezi nimi leží právě čtyři slova. Zkrácení vzdálenosti vede k prázdné množině záznamů.

Název vyhledané knihy: *1001 tipů a triků pro Visual C++ : řešení programátorských "hlavolamů", užitečné programy* / Radek Chalupa

Dotazy s operátorem vzdálenosti a s rozšířením slov

Operátory vzdálenosti v kombinaci s rozšířením někdy vyvolávají při vyhledávání v databázi NKC problémy. Kdy, a hlavně proč, je obtížné určit.

Dotazy typu Slova-Všechna pole

Modifikujeme dotaz D15 (tipů !4 c++) pravostranným rozšířením jednoho termínu (místo „tipů“ máme „tip*“). Výsledek je nulový. Tenhle výsledek nebyl očekáván a proto jej budeme analyzovat.

D20 Slova-Všechna pole= (tip* !4 c++) 0

Příčinou prázdné množiny vyhledaných záznamů není přetečení; dotazy Slova-Všechna pole= (tip*) 1259 a Slova-Všechna pole= (c++) 71 nevedou k přetečení. Obě slova jsou jistě zdrojem pro vyhledávací pole **Slova-Všechna pole**, protože pocházejí z názvu publikace. Proto kvalifikujeme neúspěšné vyhledání jako **selhání systému.** V tomto případě bylo způsobeno rozšířením slova „tip“.

Slabší dotaz se spojkou „and“ přináší kladný výsledek a stejně i modifikace dotazu obsahující vzdálenostní operátory s rozšířením termínů:

D21 Slova-Všechna pole= (tip* and c++) 1
D22 Slova-Všechna pole= (1001* %10 c++) 1
D23 Slova-Všechna pole= (1001* !10 c++) 1

Zkoušíme další dotazy obsahující vzdálenostní operátory a rozšířené termíny. Dotazem typu „and“ nejdříve určíme max. počet záznamů, které lze vůbec vyhledat.

D24 Slova-Všechna pole= (databá* and relač*) 39
D25 Slova-Všechna pole= (databáz* %5 relač*) 39
D26 Slova-Všechna pole= (databáz* %1 relač*) 37
D27 Slova-Všechna pole= (databá* %5 relač*) 0

Systém selhal v dotazu D27, i když k přetečení nedošlo. Slova-Všechna pole= (databá*) 803 a Slova-Všechna pole= (relač*) 51. V našem příkladu se vyhledávací termíny v úspěšném a neúspěšném dotazu liší jen jedním písmenem. Budeme-li testovat další dotazy tohoto typu, dojde k dalším selháním systému. Můžeme stanovit praktické pravidlo: čím je zkrácení zprava větší, tj. úvodní část slova kratší, tím větší je pravděpodobnost selhání systému.

Dotazy typu konkrétní pole

Dosud jsme používali pro zadávání dotazu jen **jedno všeobecné logické pole:** WRD alias Slova-Všechna pole. Budeme zkoušet, zda k selhání systému nedojde, když zvolíme **pole konkrétní**. U konkrétnějších polí je snazší si představit, jak se vzdálenost slov chápe a realizuje. Příkladem konkrétnějšího pole je pole Slova-Názvy.

Opravíme selhávající dotaz D20 nahrazením pole WRD polem Slova-Názvy. Ani po opravě dotazu systém nevyhledal odpovídající záznam, což hodnotíme jako selhání systému.

D28 Slova-Názvy= (tip* !4 c++) 0
http://sigma.nkp.cz/F/... ?func=find-c&ccl_term=wtl%3D%28tip*+%214+c%2B%2B%29

Opravíme druhý selhávající dotaz D27 zavedením specifického pole (wkw). Změna pomohla, vyhledání bylo správné a k selhání systému nedošlo.

D29 Klíčová slova= (databá* %5 relač*) 34
http://sigma.nkp.cz/F/... ?func=find-c&ccl_term=wkw%3D%28datab%C3%A1*+%255+rela%C4%8D*%29

Zkusíme další sadu dotazů s polem **Slova-Názvy.** Víme, že v databázi NKC existuje kniha s názvem: *Microsoft Visual Basic C++ 6.0.* Viděli jsme, že u některých částí slova „Microsoft“ k přetečení nedochází. Ověříme chování systému vzhledem k rozšíření a vzdálenostním operátorům. Stačí určit přechod mezi selháním a úspěchem. Systém někdy selhává, jindy nikoli.

D30 Slova-Názvy= (micros* !10 c++) 0
http://sigma.nkp.cz/F/... ?func=find-c&ccl_term=wtl%3D%28micros*+%2110+c%2B%2B%29

D31 Slova-Názvy= (microso* !10 c++) 2
http://sigma.nkp.cz/F/... ?func=find-c&ccl_term=wtl%3D%28microso*+%2110+c%2B%2B%29

Hledání co nejkratšího úvodního řetězce slova s pravostranným rozšířením

Úvodní řetězce slov s rozšířením jsou v případě kladného výsledku vyhledání dosti dlouhé (5 nebo 6 znaků). Pokusíme se najít slovo, které by se dalo s úspěchem rozšířit a jehož úvodní řetězec by byl dostatečně krátký. Přišli jsme na „slovo" cclxxx, které je označením čísla stránky v římské notaci. Zjistili jsme počty vyhledaných záznamů pomocí jednoslovných výrazů.

D32 Slova-Všechna pole= (ccl) 2
D33 Slova-Všechna pole= (cclxxx) 1
D34 Slova-Všechna pole= (ccl*) 7
D35 Slova-Názvy= (ccl*) 6
D36 Slova-Názvy= (reforma) 398
D37 Slova-Názvy= (střední) 0, příliš mnoho záznamů

D38 Slova-Názvy= (reforma !2000 ccl*) 1
D39 Slova-Názvy= (střední !2000 ccl*) 1

Reforma střední školy v ČSR = Secondary School Reforms in Czechoslovakia (see page CCLXXX) / Emil Čapek

Zajímavý je kombinovaný dotaz D39 se slovem „střední". Samostatné užití slova „střední" vede k výsledku „Slova-Názvy= (střední) 0, příliš mnoho záznamů". Dosud jsme vycházeli z předpokladu, že slovo kombinovaného dotazu nesmí vyvolat přetečení.

Dotazy typu blízkost podle NKC (adjacency)

Neúspěšný dotaz D27 lze do určité míry nahradit dotazem D40, který využívá opaku pojmu blízkosti z nápovědy NKC. Nepožadujeme ani závazné pořadí slov, ani max. počet vložených slov. V URL dotazu se musí objevit parametr „adjacent" s hodnotou N.

D40 Slova-Všechna pole= (databá* relač*) 39 &adjacent=N
http://sigma.nkp.cz/F/... ?func=find-c&ccl_term=datab%C3%A1*+rela%C4%8D*&adjacent=N

V dotazu D41 požadujeme závazné pořadí, nulovou vzdálenost termínů a vyhledávání ve všech polích, ale výsledkem je prázdná množina záznamů. V záznamu č. 000104048 najdeme mezi předmětovými hesly řetězec „databáze relační" (dotaz bez blízkosti), ale nejsme si jisti, zda se v D41 jedná o selhání systému či nikoli. V D42 zavedeme specifické pole a dostaneme neprázdnou množinu vyhledaných záznamů.

D41 Slova-Všechna pole= (databá* relač*) 0 &adjacent=Y
http://sigma.nkp.cz/F/... ?func=find-c&ccl_term=datab%C3%A1*+rela%C4%8D*&adjacent=Y

D42 Klíčová slova= (databá* relač*) 7
http://sigma.nkp.cz/F/... ?func=findc&ccl_term=wkw%3D%28datab%C3%A1*+rela%C4%8D*%29&adjacent=Y

Selhání v dotazech se vzdálenostním operátorem a rozšířením

Porovnejme velmi podobné neúspěšné a úspěšné dotazy.

D20 Slova-Všechna pole= (tip* !4 c++) 0
D28 Slova-Názvy= (tip* !4 c++) 0

D27 Slova-Všechna pole= (databá* %5 relač*) 0
D25 Slova-Všechna pole= (databáz* %5 relač*) 39
D29 Klíčová slova= (databá* %5 relač*) 34

D30 Slova-Názvy= (micros* !10 c++) 0
D31 Slova-Názvy= (microso* !10 c++) 2

D41 Slova-Všechna pole= (databá* relač*) 0 &adjacent=Y
D42 Klíčová slova= (databá* relač*) 7 &adjacent=Y

Úspěch se dostaví rozšířením operandu dotazu o jeden znak nebo nahrazením obecného pole polem specifičtějším, aniž změníme délku. Toto pozorování může sloužit jako vodítko pro napravování neúspěšných dotazů, ale rozhodně nikoli o přesný předpis, který funguje za všech okolností. Pokusíme se analyzovat jak zvětšení délky úvodního řetězce, tak změnu obecnosti logického pole.

Ve vyhledávacích systémech s invertovaným souborem ovlivňuje pravostranné rozšíření i obecnost logického pole rozsah zpracovávaných dat.

Mějme tato slova v invertovaném souboru odpovídající operandu dotazu „databáz*“ a „relač*“.

Databáze	Relační
Databázemi	Relačních
Databázi	Relačním
Databázový	Relačními
Databází	
Databázím	

Dotaz s operandem „*databáz**“ je logicky ekvivalentní výrazu

(Databáze or Databázemi or Databázi or Databázový or Databází or Databázím)

Obdobný ekvivalentní výraz lze vytvořit i pro druhý řetězec. Systém nahrazuje dlouhé samohlásky krátkými a proto bude výraz dále upraven.

V systému ALEPH nelze zobrazit slova se stejným úvodním řetězcem a zjistit jejich počet. Takovou funkci zabezpečuje např. CDS/ISIS. Čím kratší řetězec, tím vyšší pravděpodobnost, že logická formule s or bude delší a zpracování se prodlouží. Vzniká otázka, zda systém ALEPH nemá nějaké vnitřní omezení na počet členů v disjunkci. A dále otázka, zda nebude třeba vyčlenit větší **paměťový prostor** pro dotaz a jeho zpracování ve vnitřní paměti.

Mějme tyto názvy ve dvou různých záznamech:

MFN=5 *Databáze, jejich vytváření, aktualizace a využívání. Relační algebra a relační databáze.*

MFN=10 *Relační databáze.*

Ke klíčům „databáze“ a „relační“ invertovaného souboru obecně existuje množina odkazů. **Odkaz** se skládá z čísla záznamu (ve shodě s CDS/ISIS budeme toto číslo označovat jako MFN), z něhož je slovo vyňato, z pole, odkud bylo vyňato, z uvedení výskytu pole (opakovatelná/neopakovatelná pole) a pořadového čísla slova v poli. Obdobně existují odkazy k dalším slovům názvu. Odkazy ukazují do základního souboru záznamů.

Abychom mohli úspěšně vyhledávat pomocí základních logických spojek, stačí, aby odkaz obsahoval jen číslo MFN. Pro vyhledávání pomocí vzdálenostních operátorů musíme odkaz doplnit o další údaje.

Je zřejmé, že úplný odkaz má větší nároky na paměťový prostor než prostý odkaz na MFN. V CDS/ISIS je úplný odkaz dlouhý 8B, odkaz na MFN jen 4B, tj. úplný odkaz je dvakrát delší než prostý odkaz.

Klíč = databáze

Číslo odkazu	MFN	Číslo pole	Číslo výskytu pole	Pořadové číslo slova v poli	Poznámka
1	5	10	1	1	název
2	5	10	1	11	název
3	10	10	1	2	název

Tab. 1 - Odkazy ke klíči "databáze"

Klíč = relační

Číslo odkazu	MFN	Číslo pole	Číslo výskytu pole	Pořadové číslo slova v poli	Poznámka
1	5	10	1	1	název
2	10	10	1	1	název

Tab. 2 - Odkazy ke klíči "relační“

Jedno slovo se může v záznamu vícekrát opakovat. Při vyhledávání pomocí spojky and stačí, když vybereme jeden odkaz ze skupiny odkazů určitého pole prvního klíče a zkombinujeme ho se všemi odkazy druhého klíče.

Jestliže nezáleží na číslu pole, pak porovnáme pouze MFN. Jsou-li shodná, došlo ke shodě a záznam je relevantní k dotazu and.

Bereme-li v úvahu ještě doplňkové informace o umístění slova v poli, dostaneme obecně více porovnání. Navíc kromě MFN porovnáme další údaje v odkazu. Tím vzniká eventuální větší potřeba místa v operační paměti. Jaká bude jeho skutečná potřeba, závisí na zvolených datových strukturách a postupu při vyhodnocování dotazu.

Na základě provedeného rozboru můžeme stanovit tyto důležité **parametry vyhodnocování dotazu**:

- počet záznamů v databázi,
- počet slov v invertovaném souboru odpovídajících zprava rozšířenému úvodnímu řetězci,
- celkový počet odkazů odpovídajících všem klíčům rozšířeného úvodního řetězce,
- celkový počet odkazů odpovídajících všem klíčům rozšířeného úvodního řetězce pro zvolené pole.

V systému ALEPH může uživatel zjistit pouze počet vyhledaných záznamů. Nemůže zjistit **interval slov** odpovídající úvodnímu řetězci slova zkráceného zprava, tj. počet odpovídajících klíčů v invertovaném souboru. Setkává se s horním omezením na počet vyhledaných záznamů.

Nicméně se zdá, že počty vyhledaných dokumentů podle jednotlivých prvků dotazu ukazují, že zpřesnění operandu dotazu nebo zvýšení specifičnosti pole snižují odpovídající počet záznamů a vedou k úspěšnosti celého dotazu (a naopak). Tuto skutečnost jsme již dříve uvedli, nyní ji ilustrujeme na příkladech.

Číslo	Selhání	Úspěch	Poznámka
1	**D27 Slova-Všechna pole= (databá* %5 relač*) 0** Slova-Všechna pole= (databá*) 803 Slova-Všechna pole= (relač*) 51	D29 *Klíčová slova*= (databá* %5 relač*) 34 Klíčová slova= (databá*) 634 Klíčová slova= (relač*) 38	Specifičnost pole.
2	**D27 Slova-Všechna pole= (databá* %5 relač*) 0** Slova-Všechna pole= (databá*) 803 Slova-Všechna pole= (relač*) 51	D25 Slova-Všechna pole= (*databáz** %5 relač*) 39 Slova-Všechna pole= (databáz*) 694 Slova-Všechna pole= (relač*) 51	Delší operand.
3	Neexistuje rozumný příklad	D38 Slova-Názvy= (střední !2000 ccl*) 1 D37 Slova-Názvy= (střední) 0, příliš mnoho záznamů D35 Slova-Názvy= (ccl*) 6	Není rozumný příklad selhání. Krátký úvodní řetězec.
4	**D20 Slova-Všechna pole= (tip* !4 c++) 0** **D28 Slova-Názvy= (tip* !4 c++) 0** Slova-Názvy= (tipů) 53 Slova-Názvy= (tip*) 1040 Slova-Názvy= (c++) 66	Nepodařilo se najít úspěšnou modifikaci dotazů vlevo. Příliš krátký úvodní řetězec.	Výraz „tip*“ přináší velké množství odpovídajících záznamů.

Tab. 3 - Podrobnější porovnání úspěšných a neúspěšných dotazů

Nápověda k systému ALEPH pro databázi NKC

Obecná nápověda pro funkci vyhledávání

Je obsažena v okně „Nápověda k systému ALEPH“ se zkrácenou adresou
http://sigma.nkp.cz/F/... func=file&file_name=help-1

Obsahuje zejména tyto dílčí informace:

- práh vyhledávání,
- zavádí symboly pro rozšíření slov,
- zavádí operátory vzdálenosti se závazným nebo nezávazným pořadím slov a uvádí jen jeden příklad (česká %3 republika, resp. česká !3 republika).

Specifická nápověda pro příkazový jazyk CCL

Je umístěna do okna „NKC – Vyhledávání pomocí jazyka CCL“ se zkrácenou adresou http://sigma.nkp.cz/F/…?func=file&file_name=find-c.

Obsahuje tyto dílčí informace:
- zkratky názvů logických polí a jejich obsah,
- při vyhledávání není třeba rozlišovat velká a malá písmena,
- před slovy kombinovanými z různých polí je třeba uvádět prefix,
- zavádí symboly pro rozšíření slov,
- zavádí blízkost slov ve vztahu k formuláři CCL,
- uvnitř jednotlivých polí lze slova kombinovat pomocí logických spojek AND, OR a NOT.

Obsah obecné a specifické nápovědy je z hlediska uživatele **poněkud mlhavý**. Především není jasné, kdy lze používat operátorů vzdálenosti. Specifická nápověda pro CCL neobsahuje jediný příklad použití vzdálenostních operátorů, znamená to, že uživatel nemá tyto operátory v CCL dotazech používat? Ukázali jsme, že je často používat může, a to úspěšně; v některých případech je raději používat nemá. Navíc lze zabránit zadávání formálně nesprávných dotazů syntaktickou kontrolou spojenou s diagnostickými zprávami.

V různých systémech je možno se setkat s **příklady nesprávně vytvořených dotazů**. Uvedení příkladů formálně nesprávných dotazů lépe vymezuje rozsah správného využívání dotazovacího jazyka. Zejména tehdy, když není dotazovací jazyk představen ve formalizované podobě. Příklady nesprávných dotazů v nápovědě k databázi NKC uživatel nenajde.

V mnohých systémech se osvědčilo uvádět příklady ekvivalentních dotazů, což pomáhá uživateli k lepšímu pochopení neformálně charakterizovaného dotazovacího jazyka.

Uživatel databáze NKC není předem upozorněn na fakt, že systém nerozlišuje slova lišící se délkou samohlásky a zachází specificky s měkkými a tvrdými souhláskami při vyhledávání. Tento postup představuje jeden z možných přístupů k ukládání slov do invertovaného souboru. Druhý je naopak takový, že se odlišují i pro vyhledávání slova „vir“ a „vír“. Oba přístupy se v praxi vyhledávacích systémů vyskytují. Když nápověda obsahuje upozornění uživateli, že systém nerozlišuje velká a malá písmena, může zahrnout i chápání hlásek v systému.

V této práci jsme se zabývali spíše jednoduchými typy dotazů. Složitější dotazy v jazyce CCL mají složitější syntaxi. Ani u jednodušších dotazů jsme se nesetkali s výstižnými **diagnostickými zprávami** systému, které by umožnily uživateli dotaz syntakticky upravit. Z praxe jsou známy příklady vyhledávacích systémů, které diagnostické zprávy vydávají a pomáhají tak uživateli zvládnout často složitý dotazovací jazyk. Jestliže si systém ALEPH činí nároky na podporu vyhledávání pomocí CCL a dokonce vyhledávání v plných textech, pak by měl být uživateli více nápomocný i v otázkách syntaxe.

Závěry

Příkazový jazyk CCL implementovaný v systému ALEPH 500 v NKP představuje spolu s velkými možnostmi formátování vyhledaných záznamů silný nástroj pro úplné a přesné vyhledávání v rozsáhlé databázi NKC. Uživatel však najde pro formulaci přesných dotazů poměrně slabou podporu v obecné a specifické nápovědě, kterým chybí větší přesnost a více příkladů (i negativních). Zdá se, že u některých typů dotazů (vzdálenost a pravostranné rozšíření) může docházet k selhání, které je způsobeno krátkostí úvodního řetězce slova s pravostranným rozšířením nebo přílišnou obecností logického pole. Někdy je obtížné určit, zda selhal uživatel nebo systém.

Literatura

1. ECHO CCL-Benutzer-Handbuch. Teil 7. Befehlsübersicht, Kurzanleitung.
   Dostupný z WWW. URL: http://www.hbz-nrw.de/produkte_dienstl/fortbildung/bibliodb/um-de7.htm

2. Auwera, Vander, Y. - Bernard, L.: New proposal for a common command language in information systems. Technical Report no 118.
   Dostupný z WWW. URL: http://portal.acm.org/citation.cfm?id=984514.984518 (p51-auwera.pdf).

3. CDS/ISIS : mini-mikro (verze 2.3) : uživatelská příručka. Praha, Národní knihovna 1990. - 278 s.

4. A L E P H. Příručka pro uživatele systému. 2 OPAC.
   Dostupný z WWW. URL: http://www.sualeph.cz/a300/2_obsah.htm