Moderní metody vyhledávání dokumentů v rozsáhlých plnotextových databázích : příklad vektorového modelu

Petr Houdek[1], Josef Schwarz[2], Václav Snášel[3]

[1]Parlamentní knihovna, Sněmovní 4, Praha
houdek@psp.cz

[2]informační konzultant, Přecechtělova 2432, Praha
schwarzjv@seznam.cz

[3]Katedra informatiky, VŠB-TU Ostrava, 17. listopadu 15, Ostrava
Vaclav.Snasel@vsb.cz

Pozn.: Článek je dílčím výstupem grantového úkolu č. 201/00/1031 „*Inteligentní vyhledávání v dokumentografických informačních systémech*“ řešeného na MFF-UK za podpory Grantové agentury ČR.

Obsah

Úvod

Vývoj hardwarové i softwarové složky informačních systémů, kvalitativní i kvantitativní rozšiřování jejich datových základen a zvyšující se požadavky a nároky uživatelů představují trvalý tlak na vývoj nových modelů vyhledávání informací. V řadě systémů doposud s úspěchem funguje klasické vyhledávání založené na booleovském modelu, jež je poměrně jednoduché z hlediska principu i použití, které však již mnohde nedostačuje pro svá omezení; ta spočívají především v nemožnosti vážení dokumentů (stanovování míry jejich relevance vzhledem k rešeršnímu dotazu) a přesně určeném (binárním) způsobu vyhledávání, který často způsobuje vyhledání příliš malého nebo příliš velkého počtu dokumentů.

Zejména v souvislosti s dostupností rozsáhlých databází plných textů dokumentů se proto již delší dobu vyvíjejí a používají nové metody vyhledávání založené na statistických charakteristikách textu. Použití těchto metod vycházejících ze statistiky, lineární algebry, neuronových sítí a dalších teoretických modelů je umožněno velkým rozsahem dostupných dat, v rámci kterého se již projevují relevantní statistické vlastnosti souborů dat (dokumentů).

Obecně si lze představit vyhledávání v rozsáhlém souboru dokumentů jako zpracování bodů umístěných ve vysoce dimenzionálním prostoru. Nejdůležitějším aspektem při zpracování těchto dat je redukce dimenze prostoru provedená tak, aby vzájemný vztah jednotlivých dokumentů zůstal pokud možno zachován viz [2][3][4][6]. Statistické metody lze však pro vyhledávání použít pouze v případech, že soubor dokumentů je dostatečně rozsáhlý, v opačném případě se tento přístup ukazuje jako nevhodný (k tomu viz též [5]).

Článek, jehož hlavním účelem je popsat jednu z metod vyhledávání založenou na statistické analýze textu – vektorový model vyhledávání -, je rozdělen na tři základní kapitoly:

- kapitola *Klasifikace modelů vyhledávání* zasazuje vektorový model vyhledávání do kontextu ostatních vyhledávacích metod
- kapitola *Vektorový model vyhledávání* podrobně vysvětluje principy vektorového vyhledávání
- kapitola *Testování vektorového vyhledávání v systému AmphorA* popisuje výsledky konkrétní implementace vektorového vyhledávání, která byla v rámci řešení grantu úkolu č. 201/00/1031 „*Inteligentní vyhledávání v dokumentografických informačních systémech*“ provedena v letech 2001-2002 v databázích Kanceláře Poslanecké sněmovny Parlamentu ČR.

Klasifikace modelů vyhledávání[1]

Přestože je dnes řada informačních systémů založena na klasickém booleovském modelu vyhledávání, existuje řada dalších metod, z nichž některé jsou teprve ve fázi vývoje či dílčích implementacích, ale které jsou velmi slibné z hlediska jejich efektivity při vyhledávání ve velkých objemech dat. Mezi tyto modely patří kromě vektorového vyhledávání např. i latentní sémantické indexování, neuronové sítě, bayesovské sítě a další.

Před vlastní klasifikací modelů vyhledávání je třeba vyjasnit dvě dvojice základních termínů, které se vztahují ke způsobu, jakým jsou informace hledány, a jejich vztah k modelům vyhledávání. Jedná se o termíny vyhledávání (*retrieval*) *versus* prohlížení (*browsing*) a vyhledávání (*retrieval*) *versus* filtrace (*filtering*).

Pojmy vyhledávání a prohlížení se vztahují k různým způsobům, kterými uživatel aktivně hledá informace. Vyhledávání je založeno na definici uživatelova požadavku ve formě rešeršního dotazu formulovaného v dotazovacím jazyce daného informačního systému. Výsledkem zpracování rešeršního dotazu je množina dokumentů, která by měla být relevantní požadavku uživatele. Prohlížení na rozdíl od vyhledávání vychází z toho, že uživatel přesně nedokáže

[1] Kapitola je založena na publikaci BAEZA-YATES, R., RIBEIRO-NETO, B. *Modern information retrieval*. New York : ACM Press, 1999, s. 19-71.

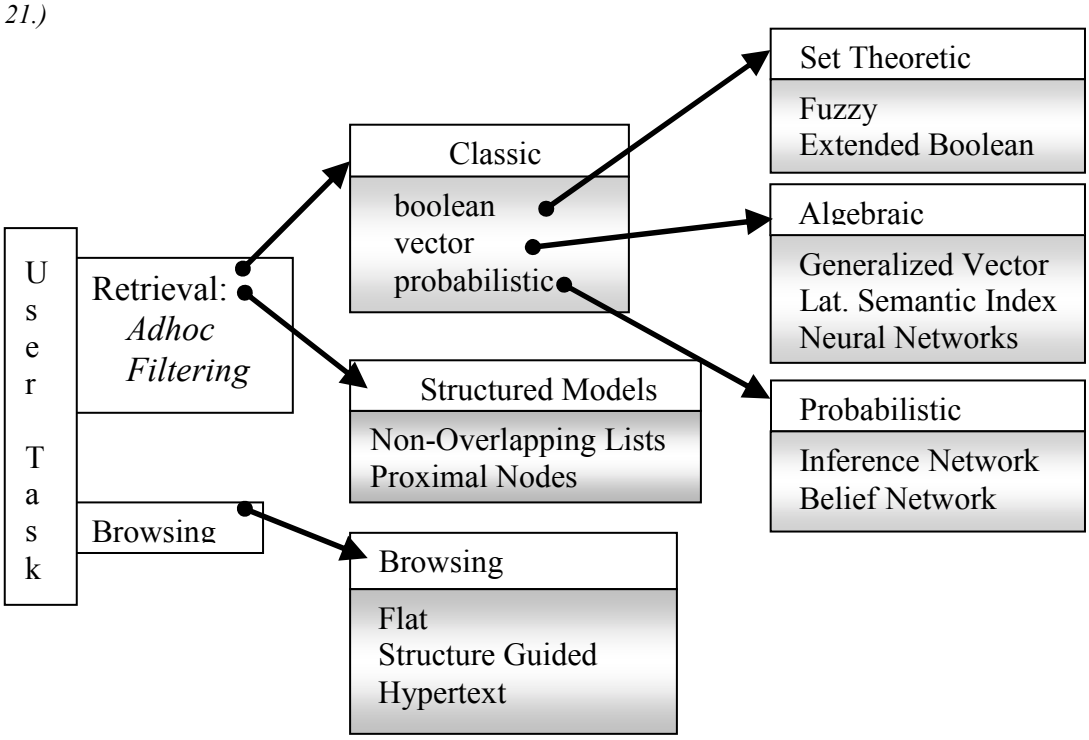
definovat svůj požadavek nebo je jeho požadavek příliš široký; pomocí interaktivního rozhraní uživatel jednoduše prochází jednotlivé dokumenty v databázi, které by mohly být relevantní pro jeho požadavek, a příp. je odkazován na další dokumenty. V průběhu prohlížení se uživatelův požadavek může změnit nebo specifikovat. Klasické vyhledávací systémy většinou pracují na principu vyhledávání, prohlížení je v informačních systémech nejčastěji zajištěno pomocí hypertextu. Moderní informační systémy obvykle kombinují vyhledávání i prohlížení, i když ve většině z nich není propojení obou principů dostatečně funkčně integrováno. Pro oba principy, vyhledávání i prohlížení, také existují různé modely vyhledávání (viz níže).

Pojmy vyhledávání a filtrace se vztahují k tomu, jakým způsobem jsou informace vyhledávány. Zatímco u vyhledávání[2] dochází na základě jednorázového uživatelova požadavku k formulaci rešeršního dotazu, pomocí kterého jsou vyhledány potenciálně relevantní dokumenty, při filtraci je definován profil uživatele, který odpovídá jeho víceméně stalému požadavku a který je používán pro obvykle periodické dodávání nových dokumentů uživateli. Modely vyhledávání se pro oba způsoby, vyhledávání a filtraci, neliší.

Modely vyhledávání lze rozdělit na klasické, mezi které patří booleovské, vektorové a pravděpodobnostní vyhledávání, a strukturované modely (modely pro vyhledávání ve strukturovaném textu), kam lze zařadit model nepřekrývajících se seznamů (non-overlapping lists) a sousedních uzlů (proximal nodes). Kromě toho existuje řada alternativ ke klasickým modelům vyhledávání, mezi které lze zařadit rozšířený booleovský model, fuzzy vyhledávání, zobecněné vektorové vyhledávání, latentní sémantické indexování, neuronové sítě, bayesovské a inferenční sítě. Speciální modely založené na hypertextovém, plošném nebo strukturovaném přístupu jsou uplatňovány v rámci systémů založeným na prohlížení. Souhrnnou klasifikaci modelů podává níže uvedený obrázek.

Obrázek č. 1 Klasifikace modelů vyhledávání

(převzato z BAEZA-YATES, R., RIBEIRO-NETO, B. Modern information retrieval. New York : ACM Press, 1999, s. 21.)



V následující kapitole je podrobněji popsán vektorový model vyhledávání.

[2] Někdy se v kontrastu s filtrací označuje jako *ad hoc* vyhledávání (*ad hoc retrieval*).

Vektorový model vyhledávání

Hlavním rysem vektorového modelu je reprezentace dokumentů a uživatelských dotazů pomocí vektorů. K objasnění principu vektorového vyhledávání slouží následující příklad:

Představme si, že se pohybujeme v databázi, kde lze vyhledávat podle této množiny klíčových slov:

databáze

relevance

koeficient přesnosti

koeficient úplnosti

zpracování textu

Nechť databáze obsahuje tři dokumenty obsahující uvedená klíčová slova:

Dokument D1: databáze, relevance, koeficient přesnosti

Dokument D2: koeficient přesnosti, koeficient úplnosti

Dokument D3: databáze, zpracování textu

Reprezentace těchto dokumentů pomocí vektorů je:

D1: (1,1,1,0,0)

D2: (0,0,1,1,0)

D3: (1,0,0,0,1)

kde 1, resp. 0 sděluje fakt, zda se dané klíčové slovo vyskytuje v dokumentu. Uvažujme dotaz na dokumenty s klíčovými slovy *databáze* a *zpracování textu*. Jeho vektor bude (1,0,0,0,1). Vyhledáme-li dokument pomocí booleovského výrazu, zjistíme, že dokument D3 je hitem (relevantním záznamem). Zajímavý však může být i dokument D1 obsahující klíčové slovo *databáze*, dokument D2 je zřejmě nerelevantní. Tato fakta lze ale též zjistit ekvivalentním způsobem, vynásobíme-li vektor dotazu skalárně s vektory dokumentů, tj. např. $(1,1,1,0,0) \times (1,0,0,0,1) = 1$. Postupnou aplikací součinu obdržíme čísla 1, 0, 2, podle kterých lze seřadit dokumenty do pořadí *D3*, *D1*, *D2*.

Pokud budeme zaznamenávat pro jednotlivé dokumenty počty výskytů jejich klíčových slov, získáme jednoduchý *systém vážení*. Např. pro dokument *D1* by mohl vypadat vektor vah jako (3,0,0,0,2). Aplikací skalárního součinu bychom obdrželi číslo 5. Je vidět, že tato čísla již nevyjadřují jednoduše shodu klíčového slova dotazu s klíčovým slovem v dokumentu, ale určité kvantitativní ohodnocení míry podobnosti dokumentu s dotazem. Pro relevanci by totiž mohlo být významné, že se dané klíčové slovo vyskytuje v dokumentu častěji než jiné klíčové slovo. Tato idea je základem vektorového modelu.

Předpokládejme, že pro indexaci všech dokumentů v databázi bylo použito celkem n různých klíčových slov $t_1 ... t_n$; potom každý dokument D_i ze souboru dokumentů **D** je reprezentován vektorem

$$D_i = (w_{i1}, w_{i2}, ..., w_{in}), \text{ kde } w_{ij} \in R$$

kde w_{ij} jsou váhy náležející klíčovému slovu t_j v identifikaci dokumentu D_i. Váha w_{ij} ohodnocuje důležitost jednotlivých klíčových slov pro identifikaci dokumentu. Váha rovna nule představuje nejnižší důležitost, váha rovna jedné důležitost nejvyšší.

Soubor dokumentů **D** je ve vektorovém modelu popsán maticí

$$\mathbf{D} = \begin{matrix} w_{11} & w_{12} & \dots & w_{1n} \\ w_{21} & w_{22} & \dots & w_{2n} \\ \dots & & & \\ \dots & & & \\ w_{m1} & w_{m2} & \dots & w_{mn} \end{matrix}$$

ve které *i*-tý řádek odpovídá *i*-tému dokumentu, *j*-tý sloupec *j*-tému klíčovému slovu.

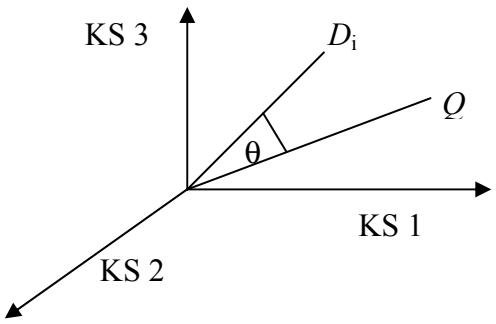
Výraz dotazu Q ve vektorovém modelu je možné formulovat jako *n*-místný vektor vah

$Q = (q_1, q_2, ..., q_n)$, kde $q_j \in R$.

Na základě dotazu Q lze pro každý dokument D_i spočítat tzv. *koeficient podobnosti* (*similarity rate*). Tento koeficient si lze představit jako "podobnost" vektoru dokumentu s vektorem dotazu ve vektorovém prostoru R^n. Pro výpočet podobnosti dokumentu D_i vzhledem k dotazu Q se používá řada vzorců. V nejjednodušším případě může být koeficient podobnosti definován vztahem

$$Sim(Q,D_i) = \frac{\sum_{k=1}^{n} (q_k \cdot w_{ik})}{\sqrt{\sum_{k=1}^{n} (w_{ik})^2 \cdot \sum_{k=1}^{n} (q_k)^2}} \qquad (*)$$

Tento koeficient podobnosti se nazývá *kosinová míra*. Kosinová míra nám pomůže získat geometrickou představu vzdálenosti mezi dvěma vektory. Na obrázku č. 2 jsou zobrazeny vektor dotazu a vektor dokumentu v prostoru tří klíčových slov (KS - třírozměrný prostor). Škály v jednotlivých osách reprezentují váhy. Kosinovou mírou se počítá kosinus úhlu, který svírá vektor dotazu s vektorem dokumentu. Čím je tato míra větší, tím je úhel menší, tj. tím „bližší“ a tedy více podobné jsou oba vektory. Děliteli ve vzorci (*) se říká také *normalizační faktor*, který stírá vliv délky dokumentu na hodnotu koeficientu *Sim*. Z toho vyplývá, že dokumenty D' a D s klíčovými slovy a, b, c, které v D mají četnosti (1,1,1) a v D' četnosti (2,2,2), povedou ke stejné hodnotě *Sim*. Neplatí totiž, že klíčové slovo, které se vyskytuje v dokumentu dvakrát, je také dvakrát důležitější.



Obrázek č. 2 Dotaz a dokument v třírozměrném prostoru

Výběr klíčových slov

Většina automatických způsobů indexování je založena na pozorování, že významnost klíčových slov pro indexování přímo souvisí s frekvencí výskytu klíčových slov v dokumentu. V praxi obvykle nejde přímo o klíčová slova, ale o jejich kmeny. Podobně je tomu i u klíčových slov dotazu.

Frekvence klíčového slova v dokumentu je číslo, které udává počet výskytů klíčové slova v dokumentu. Frekvenci klíčového slova t_j *v dokumentu* D_i označíme symbolem TF_{ij}. Někdy se používá *normalizovaná frekvence klíčového slova* t_j *v dokumentu* D_i (*NTF*) definovaná jako

$$D_i\,(NTF) = \frac{\frac{TF_{ij}}{\max_k TF_{ik}} + 1}{2}$$

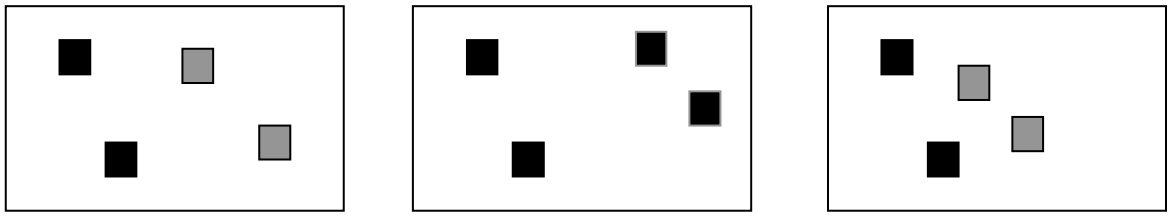
Samotná *TF* ještě nezajišťuje kvalitní výsledky informačního systému. Je nutné vzít do úvahy také frekvenci, s jakou se dané klíčové slovo vyskytuje v celé kolekci dokumentů, neboť nejvhodnějšími klíčová slova jsou ta, která se vyskytují zhruba v polovině všech dokumentů. Naopak, nevhodná jsou obecná slova, která se vyskytují ve většině dokumentů. Taková je třeba eliminovat, popř. se frekvence klíčového slova v dokumentu násobí opravnými koeficienty.

Jedna z používaných metod využívá tzv. *inverzní frekvenci klíčového slova v dokumentech* (*IDF*). *IDF* klesá se zvyšujícím se počtem dokumentů, ke kterým je dané klíčové slovo přiřazeno. *IDF* pro klíčové slovo t_j je definována předpisem

$$IDF_j = \log\left(\frac{m}{k_j}\right) + 1$$

kde m je celkový počet dokumentů v kolekci a k_j je počet dokumentů, ke kterým je přiřazeno klíčové slovo t_j. Veličině k_j se také říká *frekvence* t_j *v dokumentech* (DF_j), tedy *IDF* je skutečně inverzní vzhledem k *DF*. V krajních polohách se *IDF* chová v souladu s intuicí. Když se klíčové slovo vyskytuje ve všech dokumentech, pak log(1) = 0. V takovém případě není možné rozlišit relevantní a nerelevantní dokumenty a je tedy smysluplné zařadit dané klíčové slovo mezi nevýznamová slova. V opačném extrému, kdy se klíčové slovo vyskytuje pouze v jednom dokumentu, pak $IDF = \log m$, tj. např. pro $m = 10$ je $IDF = 1$, pro $m = 10\,000$ je $IDF = 4$ atd.

Účelem klíčových slov použitých pro indexaci je, aby rozlišily jeden dokument od druhého. Zobrazíme-li prostor dokumentů do dvourozměrného prostoru, pak přidání dobrého klíčového slova znamená, že dokumenty budou rozloženy tak, že zůstanou zachovány jejich relativní vzdálenosti. Obrázek 3 ukazuje situaci, jak se může změnit rozložení dokumentů po přidání dobrého, resp. špatného klíčového slova.



původní prostor | po přidání dobrého KS | po přidání špatného KS

Obrázek č. 3 Rozlišovací schopnost klíčových slov

Rozložení dobrých a špatných klíčových slov zvolených pro popis dokumentu ukazuje následující příklad.

Z tabulky je patrné, že zařazení klíčového slova *a* povede k tomu, že podobnost jakýchkoliv dvou dokumentů bude větší než nula, což není dobré z hlediska rozlišení dokumentů. Zahrnutí klíčového slova *r* znamená větší separaci dokumentů D_1 a D_2 od dokumentů D_3 a D_4.

Dokument	všechny KS	možný výběr	špatný výběr	dobrý výběr
D_1	*a b c d r*	*b c d*	*a b c d*	*b c d r*
D_2	*a b a d r*	*b a d*	*a b a d*	*b a d r*
D_3	*a m p q*	*m p q*	*a m p q*	*m p q*
D_4	*a x p q*	*x p q*	*a x p q*	*x p q*

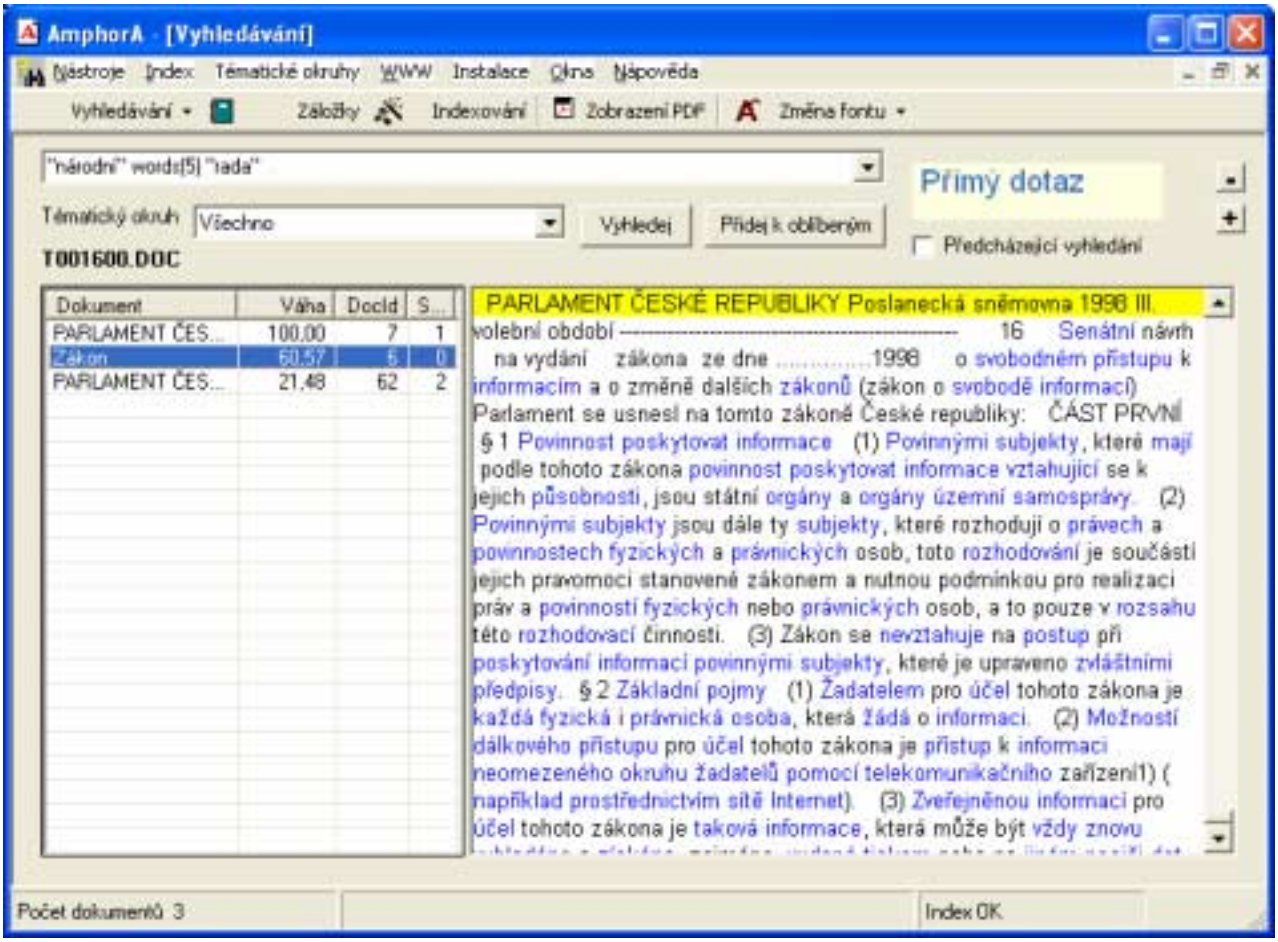
Přidání klíčového slova znamená zvýšení počtu dimenzí ve vektorovém prostoru. Dokumenty neobsahující dané klíčové slovo se nezmění. Ty, co jej obsahují, budou „odtaženy“ od těch, které jej neobsahují.

Je zřejmé, že výběr klíčových slov je jistou heuristikou pro redukci dimenze. Jako překvapivé se jeví výsledky použití náhodné redukce dimenze (viz [4] a [6]), které ukazují, že náhodně zvolená redukce se bude chovat „dobře“, to znamená, bude zachovávat vzájemné rozložení dokumentů. Experimenty uvedené v [4] naznačují, že tato metoda má velké využití v kombinaci s klasickými metodami jako je například faktorová analýza viz [13].

Jako nejpokročilejší metoda se v současné době jeví latentní sémantické indexování, kde se redukce dimenze provádí pomocí metody SVD (singular value decomposition) (viz [2][3][9]). Experimenty prováděné v rámci grantu ukazují, že jako vhodné se jeví redukovat dimenzi na několik stovek. Výsledky ukazují na zajímavou paralelu s experimenty prováděnými v 80. letech V. Smetáčkem, které prokázalo, že sémantické vyhledávání lze založit na několika stech sémů (základních sémantických jednotek) (viz blíže [10] a [11])

Testování vektorového vyhledávání v systému AmphorA

V rámci řešení grantového úkolu č. 201/00/1031 *„Inteligentní vyhledávání v dokumentografických informačních systémech“* byl implementován vektorový model vyhledávání v testovací databázi AmphorA, jejíž datovou základnu poskytla Kancelář Poslanecké sněmovny Parlamentu ČR. Testovací databáze obsahovala 30000 dokumentů a 270000 různých slov, což lze chápat jako 30000 bodů v 270000-dimenzionálním prostoru. Pomocí výše uvedených heuristických postupů bylo možné při vyhodnocování dotazu zredukovat tuto dimenzi na několik set souřadnic. To vedlo k poměrně dobré efektivitě při vyhodnocování dotazu; čas potřebný pro vyhledání v testovací databázi se pohyboval okolo 10 s. Následující obrázek ukazuje uživatelské rozhraní systému AmphorA. Vektorový dotaz je zde možno kombinovat s booleovským dotazem. Na základě vyhodnocení booleovského dotazu získáme kolekci dokumentů, ze které se pak vybírá dokument sloužící jako dotaz ve vektorovém dotazu. Podobnost dokumentů je měřena pomocí kosinové míry. Výsledek je v uživatelském rozhraní prezentován jako procento podobnosti dokumentů to znamená $Sim(Q,D)\cdot 100$.



Popis testovacích dat

Testování vektorového vyhledávání bylo provedeno na datech, která jsou součástí *Digitální knihovny Český Parlament* (http://www.psp.cz/eknih/). Digitální knihovna je společným projektem Parlamentní knihovny a Odboru informatiky Kanceláře Poslanecké sněmovny Parlamentu ČR. Cílem databáze je zpřístupňovat informace o současné i minulé činnosti českého parlamentu nejširší veřejnosti prostřednictvím internetu. O projektu již byla publikována řada článků (viz např. [12], [14], [15] nebo [16]), základní informace je rovněž možné získat na stránkách samotné Digitální knihovny, pro účely tohoto článku bude proto popis Digitální knihovny omezen pouze na její dílčí vlastnosti, které jsou podstatné z hlediska složení a struktury testovací databáze.

Obsah databáze a typy dokumentů

Digitální knihovna obsahuje tzv. **parlamentária,** tj. dokumenty, které vznikají při činnosti parlamentu. Jedná se především o dvě základní skupiny dokumentů:

1. **těsnopisecké zprávy** neboli **stenoprotokoly** (tj. doslovné záznamy ze všech jednání parlamentu) - Digitální knihovna obsahuje kompletní plné texty těsnopiseckých zpráv od roku 1908 do současnosti
2. **sněmovní tisky** (materiály předkládané k projednání na schůzích parlamentu) - Digitální knihovna obsahuje plné texty sněmovních tisků od roku 1918 až do současnosti zhruba v tříčtvrtinovém pokrytí (ke 100% pokrytí dojde v polovině roku 2003).

Novější volební období Poslanecké sněmovny (počínaje 1. volebním obdobím Poslanecké sněmovny, které začalo v lednu 1993) jsou navíc doplněna o **další dokumenty**, např. usnesení

Poslanecké sněmovny, usnesení jednotlivých výborů, zprávy výborů, zápisy z jednání výborů, pozvánky na schůze atd.

Těsnopisecké zprávy novějších volebních období jsou také doplněny o **statistiky hlasování**.

Testování vektorového vyhledávání bylo provedeno na databázi sněmovních tisků (http://www.psp.cz/sqw/sntisk.sqw), ve které se vyskytují následující **typy tisků**:

- **návrhy zákonů a novel zákonů** (jejich součástí je důvodová zpráva)
- **mezinárodní smlouvy**
  - bilaterální
  - multilaterální
- **státní rozpočty**
- **písemné interpelace** a **odpovědi na písemné interpelace**
- **zprávy**
- **stanoviska vlády**
- **usnesení výborů**
- **pozměňovací návrhy**
- další dokumenty - např. oponentní zprávy

Formát a struktura dokumentů

V Digitální knihovně se používají dva formáty dokumentů – HTML a MS Word 97. Dokumenty ve formát HTML převažují, formát MS Word 97 je originálním formátem sněmovních tisků 3. a 4. (současného) volebního období Poslanecké sněmovny, které jsou rovněž dostupné ve formátu HTML.

Digitální knihovna je strukturována chronologicky podle jednotlivých volebních období, která jsou dále členěna podle typů dokumentů.

Dokumenty Digitální knihovny nejsou strukturovány a neobsahují žádné metadatové záznamy. Řešením tohoto problému se zabývá projekt *Oběh dokumentů mezi ústředními orgány státní správy* (http://www.senat.cz/ISO-8859-2.cgi/info/navrhzak/index.htm), jehož dlouhodobějším cílem je zajistit tvorbu a vzájemnou výměnu dokumentů ve formátu XML.

Výběr dat pro testování vektorového vyhledávání

Pro testování vektorového vyhledávání byly zvoleny sněmovní tisky 1. až 3. volebního období Poslanecké sněmovny (tj. od ledna roku 1993 do června roku 2002). Z testování byly vyřazeny stenoprotokoly, a to z těchto dvou základních důvodů:

1. Stenoprotokoly jsou rozděleny do malých souborů po 10 minutách, aby je bylo možné co nejrychleji zveřejnit na Internetu, celá rozprava k danému tématu je tak rozčleněna na několik částí; tematické shlukování takto segmentovaných dokumentů by bylo velmi obtížné.
2. Stenoprotokoly jsou doslovným záznamem jednání Poslanecké sněmovny a mají proto z hlediska tematické struktury textu značně odlišnou kvalitu od sněmovních tisků. Jejich zařazením do testovací databáze by byla do dat vnesena inkonsistence, která by mohla negativně ovlivnit výsledky testování.

Metodika testování

Popis standardních testovacích metodik

Standardní metodiky testování vyhledávacích technik jsou založeny na měření efektivity vyhledávání pomocí koeficientů úplnosti (R) a přesnosti (P). Koeficient úplnosti obvykle v procentech vyjadřuje, kolik relevantních dokumentů z jejich celkového počtu v databázi bylo nalezeno, koeficient přesnosti vyjadřuje procentní podíl relevantních dokumentů v množině všech vyhledaných dokumentů. Matematické vyjádření koeficientů úplnosti a přesnosti je následující:

$$R=\frac{a}{a+b} \qquad P=\frac{a}{a+c}$$

kde
a = počet nalezených relevantních dokumentů
b = počet nenalezených relevantních dokumentů
c = počet nalezených nerelevantních dokumentů

Vzhledem k tomu, že celkové míry úplnosti a přesnosti nemají dostatečnou vypovídací hodnotu, používá se metodika založená na sledování přesnosti v jedenácti normalizovaných úrovní úplnosti (*precision at 11 standard recall levels – blíže viz [1]*). 11 normalizovaných úrovní úplností je představováno hodnotami 0, 10, 20, 30, 40, 50, 60, 70, 80, 90 a 100 % úplnosti s tím, že ke každé úrovni úplnosti je uvedena odpovídající míra přesnosti.

Výsledky jednotlivých rešerší je třeba obvykle normalizovat (popř. interpolovat) na příslušných 11 úrovní úplnosti; normalizovaná přesnost na dané úrovni úplnosti (např. 20%) je rovna maximální přesnosti z množiny výsledků vztažených k dané úrovni úplnosti a následující úrovni úplnosti (tj. 20-30%).

Příklad:

Hodnoty R a P:

R	***P***
9,1	100
12,1	57,1
15,2	62,5
18,2	66,7
21,2	70
24,2	66,7
27,3	69,2
30,3	71,4

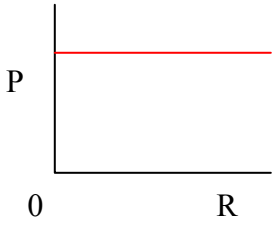
budou normalizovány takto:

R	***P***
0	100
10	66,7
20	70
30	71,4

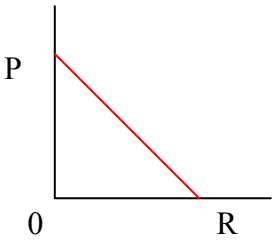
Normalizace hodnot úplnosti a přesnosti je podstatná zejména při srovnávání výsledků z více než jednoho vyhledávání.

Metodika sledování přesnosti v jedenácti normalizovaných úrovní úplnosti umožňuje v podstatě spojitě sledovat, jak se mění přesnost v závislosti na úplnosti v průběhu vyhledávání (resp. v průběhu prohlížení výsledků uživatelem).

V ideálním případě by při zvyšující se úplnosti zůstávala přesnost stále 100%, grafické vyjádření tohoto stavu by vypadalo takto:



Tento stav je ovšem v praxi nedosažitelný, empiricky bylo zjištěno, že přesnost je nepřímo úměrná úplnost (při zvyšující se úplnosti klesá přesnost vyhledávání a naopak) a grafické vyjádření má proto tento průběh:



Kromě sledování přesnosti v jedenácti normalizovaných úrovní úplnosti byly pro hodnocení výsledků vyhledávání použity také následující míry:

Harmonický průměr úplnosti a přesnosti

Jedná se o jednoduchou míru kombinující přesnost a úplnost. Hodnota výsledku je závislá na obou veličinách, vysoká hodnota harmonického průměru znamená, že je vysoká úplnost i přesnost. Harmonický průměr se vypočte jako:

$$F = \frac{2}{\frac{1}{r_j} + \frac{1}{p_j}}$$

kde:

r_j = hodnota R pro j-tý dokument v pořadí

p_j = hodnota P pro j-tý dokument v pořadí

Ohodnocený průměr úplnosti a přesnosti

Míra, která zohledňuje důležitost úplnosti nebo přesnosti pro uživatele; matematické vyjádření míry je:

$$E = \frac{1 + b^2}{\frac{b^2}{r_j} + \frac{1}{p_j}}$$

kde:

r_j = hodnota R pro j-tý dokument v pořadí

p_j = hodnota P pro j-tý dokument v pořadí

b je parametr, který zohledňuje důležitost přesnosti nebo úplnosti z hlediska uživatele. Pokud je b menší než 1, uživatel upřednostňuje úplnost před přesností, pokud je b větší než 1, uživatel preferuje přesnost před úplností.

Konkrétní postup při testování

Testování vektorového vyhledávání v systému AmphorA bylo založeno na srovnání dvou množin dokumentů a generování hodnot pro výpočet standardních charakteristik (viz výše). Základem bylo vytvoření referenční množiny dokumentů (RMD), která byla srovnávána s testovací množinou dokumentů (TMD) získanou na základě výsledků vyhledávání pomocí vektorového vyhledávání. Vzhledem k tomu, že vektorové vyhledávání probíhalo na třech úrovních podobnosti dokumentů, jedné RMD vždy odpovídaly tři TMD.

Vytvoření RMD a TMD a jejich srovnání bylo provedeno tímto způsobem:

Vytvoření RMD

Z testovací databáze byl vybrán referenční dokument X[3], k němuž byly pomocí vyhledávacích technik dostupných v uživatelském rozhraní Digitální knihovny (vyhledávání pomocí klíčových slov z názvů dokumentů a rejstříku ke sněmovním tiskům) nalezeny obsahově podobné dokumenty[4]. Tato množina dokumentů byla pro účely testování považována za 100% úplnou a 100% relevantní množinu dokumentů. Pro účely výpočtu standardních charakteristik byla zaznamenána čísla jednotlivých dokumentů a celkový počet dokumentů v referenční množině.

Vytvoření TMD

V testovací databázi v systému Amphora byl nalezen dokument X, který byl použit jako základ pro vektorové vyhledávání. Vektorové vyhledávání bylo provedeno pro každou RMD vždy třikrát pomocí vyhledávání podle příkladu (*query by example*), a to se třemi hodnotami atributu pro podobnost dokumentů (*threshold*): 0,700, 0,500 a 0,050. Tak vznikly tři TMD (TMD0700, TMD0500 a TMD0050), k nimž byla zaznamenána čísla jednotlivých dokumentů a počet vyhledaných dokumentů.

Komparace TMD a RMD

Po vytvoření TMD a RMD byla provedena jejich komparace, tzn. že v TMD byly zjištěny relevantní dokumenty (vzhledem k RMD), jejich počet a byly vypočteny hodnoty koeficientů úplnosti (R) a přesnosti (R) jako podklad pro další analýzy. Výsledkem komparace TMD a RMD byla tabulka, jež obsahovala tyto údaje:

Číslo referenčního dokumentu				
Hodnota atributu pro podobnost dokumentů (*threshold*)				
Pořadí relevantních dokumentů v TMD	Počet relevantních dokumentů v TMD (vzhledem k RMD)	Počet dokumentů v RMD	Úplnost (R) [%]	Přesnost (P) [%]

Příklad výsledku komparace TMD a RMD:

Referenční tisk 1998PS\TISKY\T0335			
Treshold 0.700			

[3] Jednalo se vždy o návrh zákona, návrh novely zákona, mezinárodní smlouvu nebo státní rozpočet. Dokumenty typu písemné interpelace, stanoviska vlády, usnesení výborů a pozměňovací návrhy nebyly do RMD zařazeny; důvodem u většiny z nich byl jejich druhotný charakter vzhledem k první množině dokumentů (návrhy zákonů ad.) nebo nedostupnost v plném textu. Referenčním dokumentem se mohl stát pouze sněmovní tisk z 3. volebního období.

[4] Obvykle se jednalo např. o novely příslušného zákona, skupinu zákonů spojených společnou oblastí právní úpravy (např. silniční doprava) nebo určité typy mezinárodních smluv (např. o zamezení dvojího zdanění) apod.

Pořadí relevantních dokumentů v TMD	Počet relevantních dokumentů v TMD (vzhledem k RMD)	Počet dokumentů v RMD	Úplnost (R) [%]	Přesnost (P) [%]
1	1	15	6,7	100,0
3	2	15	13,3	66,7
4	3	15	20,0	75,0
5	4	15	26,7	80,0
9	5	15	33,3	55,6
11	6	15	40,0	54,5
12	7	15	46,7	58,3
13	8	15	53,3	61,5
19	9	15	60,0	47,4
29	10	15	66,7	34,5
476	11	15	73,3	2,3
654	12	15	80,0	1,8

Z tabulky je patrné, že bylo vyhledáno pouze 12 z 15 relevantních dokumentů (tzn. že max. úplnost je 80%), 10. relevantní dokument byl vyhledán jako 29. ze všech dokumentů v TMD (tj. přesnost je 34,5), 12. relevantní dokument se ocitl až na 654. pozici, tzn. že přesnost na 12. úrovni úplnosti je pouze 1,8%.

Normalizace výsledků komparace TMD a RMD

Po stanovení základních údajů bylo třeba provést normalizaci úplnosti a přesnosti na 11 úrovní úplnosti na základě výše uvedené metodiky (viz kap. *Popis standardních testovacích metodik*). Normalizované výsledky pak již obsahovaly pouze příslušné hodnoty úplnosti a přesnosti pro danou úroveň podobnosti dokumentů (*threshold - TH*), číslo referenčního dokumentu a počet dokumentů v RMD

Příklad normalizovaných hodnot R a P (na základě tabulky v předchozím příkladě):

TH 0,700

Referenční tisk č. 335

Počet dokumentů v RMD: 15

Standardní úroveň úplnosti [%]	Normalizovaná hodnota přesnosti [%]
0	100,0
10	66,7
20	80,0
30	55,6
40	58,3
50	61,5
60	47,4
70	2,3
80	1,8
90	0,0
100	0,0

Výsledky testování a jejich interpretace a hodnocení

V rámci testovacího vyhledávání bylo vytvořeno 72 RMD obsahujících celkem 615 dokumentů a vyhledáno 216 TMD (vždy 72 TMD pro TH 0,700, 0,500 a 0,050). Vzhledem

k celkovému objemu dokumentů v testovací databázi (30000 dokumentů) není počet RMD velký, nicméně orientační údaje o testovacích datech poskytuje.

Pro výslednou analýzu byl pro jednotlivé úrovně úplnosti stanoven průměr a medián odpovídajících hodnot koeficientu přesnosti. Vzhledem k tomu, že TMD obsahovaly extrémy zahrnující nízké i vysoké hodnoty,[5] zdá se, že reálnému stavu více odpovídají hodnoty mediánu, nicméně v dalších přehledech jsou pro úplnost zařazeny obě hodnoty (průměr i medián).

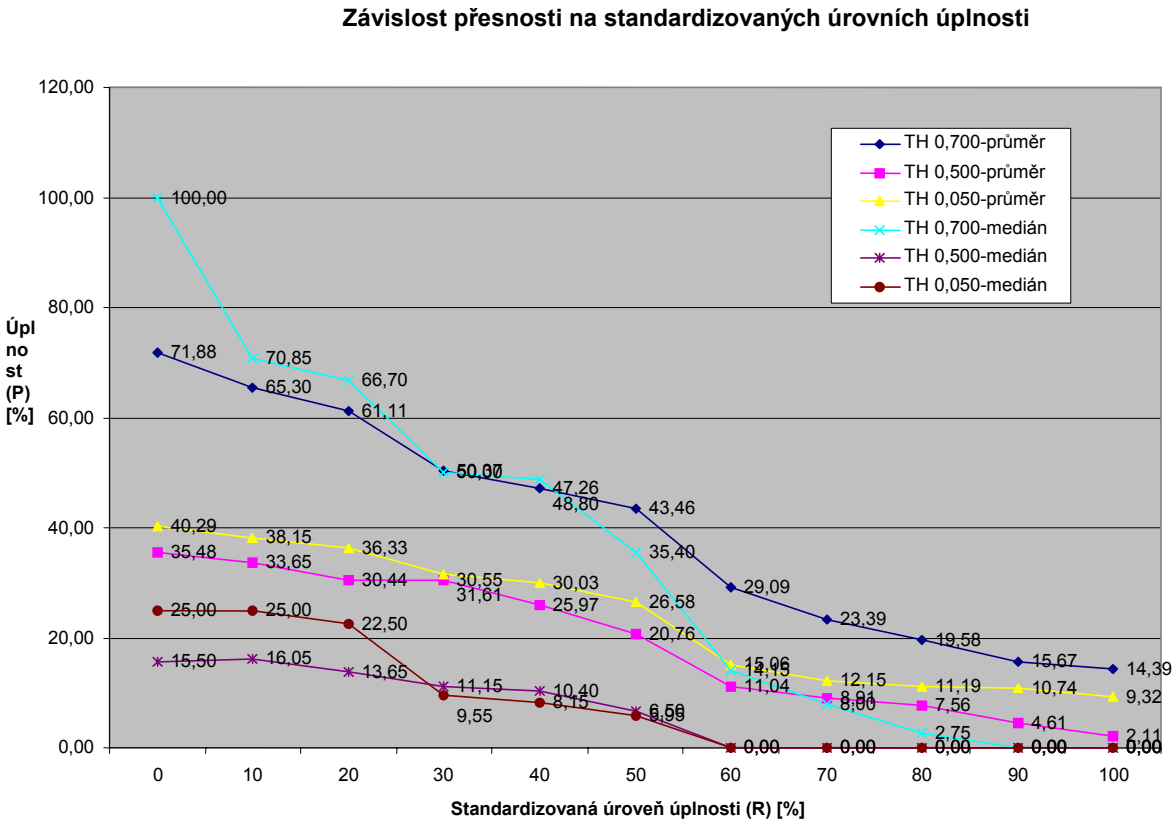
Sledování přesnosti na základě standardizovaných úrovní úplnosti:

Tabulka č. 1 Závislost přesnosti na standardizovaných úrovních úplnosti

Standardní úroveň úplnosti [%]	*Normalizovaná hodnota přesnosti - průměr [%]*			*Normalizovaná hodnota přesnosti – medián [%]*		
	TH0700	*TH0500*	*TH0050*	*TH0700*	*TH0500*	*TH0050*
0	71,88	35,48	40,29	100,00	15,50	25,00
10	65,30	33,65	38,15	70,85	16,05	25,00
20	61,11	30,44	36,33	66,70	13,65	22,50
30	50,37	30,55	31,61	50,00	11,15	9,55
40	47,26	25,97	30,03	48,80	10,40	8,15
50	43,46	20,76	26,58	35,40	6,50	5,95
60	29,09	11,04	15,06	14,15	0,00	0,00
70	23,39	8,91	12,15	8,00	0,00	0,00
80	19,58	7,56	11,19	2,75	0,00	0,00
90	15,67	4,61	10,74	0,00	0,00	0,00
100	14,39	2,11	9,32	0,00	0,00	0,00

[5] Tyto extrémy byly způsobeny zejména RMD, které obsahovaly pouze 2 dokumenty, nebo prázdnými TMD (tj. TMD, které neobsahovaly žádný dokument).

Graf č. 1



Z dat zobrazených v tabulce a grafu vyplývá, že:

- k nejefektivnější vyhledávání dokumentů dochází, pokud systém pracuje s parametrem podobnosti dokumentů TH 0,700, ostatní dvě hodnoty parametru (TH 0,500 a TH 0,050) jsou podstatně méně úspěšné z hlediska přesnosti i z hlediska úplnosti (pokud bereme v úvahu výsledky dle mediánu – pomocí obou parametrů je dosaženo méně než 60% celkové úplnosti)
- nejúspěšnější vyhledávání s parametrem TH 0,700 dosahuje při 30% úplnosti již jen 50% přesnosti (výsledky na základě mediánu i průměru), přičemž s větší než 50% úplností se tento podíl podstatně snižuje – 15% odpovídá 60% úplnosti, 8% přesnosti odpovídá 70% úplnosti atd. (podle mediánu).
- Celková dosažená úplnost je podle mediánu nižší než 90% (100% úplnost bylo dosaženo v 26 případech ze 72 s průměrnou přesností 14%, 90% úplnosti bylo dosaženo v 30 případech, 80% úplnost bylo dosaženo v 39, tj. více než polovině případů).

Otázky týkajících se nízkých hodnot úplnosti a přesnosti (zvláště dle mediánu) lze částečně zdůvodnit povahou testovacích dat:

1. úplnost mohla být negativně ovlivněna zejména tím, že některé dokumenty neobsahují text použitelný při plnotextovém vyhledávání, např. novely zákonů někdy obsahují minimum textu, vztahujícímu se k novelizovanému zákonu; přestože byl takový dokument zařazen do RMD, nebyl již vyhledán v TMD. Je však třeba uvést, že tento nedostatek se zapříčiněn spíše specifickou sémantickou strukturou dokumentů a negativně by ovlivňoval libovolnou techniku plnotextového vyhledávání.
2. přesnost je závislá na počtu vyhledaných nerelevantních dokumentů, jako relevantní byly stanoveny dokumenty uvedené v RMD; vzhledem ke způsobu výběru dokumentů do

RMD (viz výše) se mohlo stát, že do TMD byly vyhledány další dokumenty, které nebyly relevantní z hlediska RMD, ale byly relevantní z hlediska dotazu

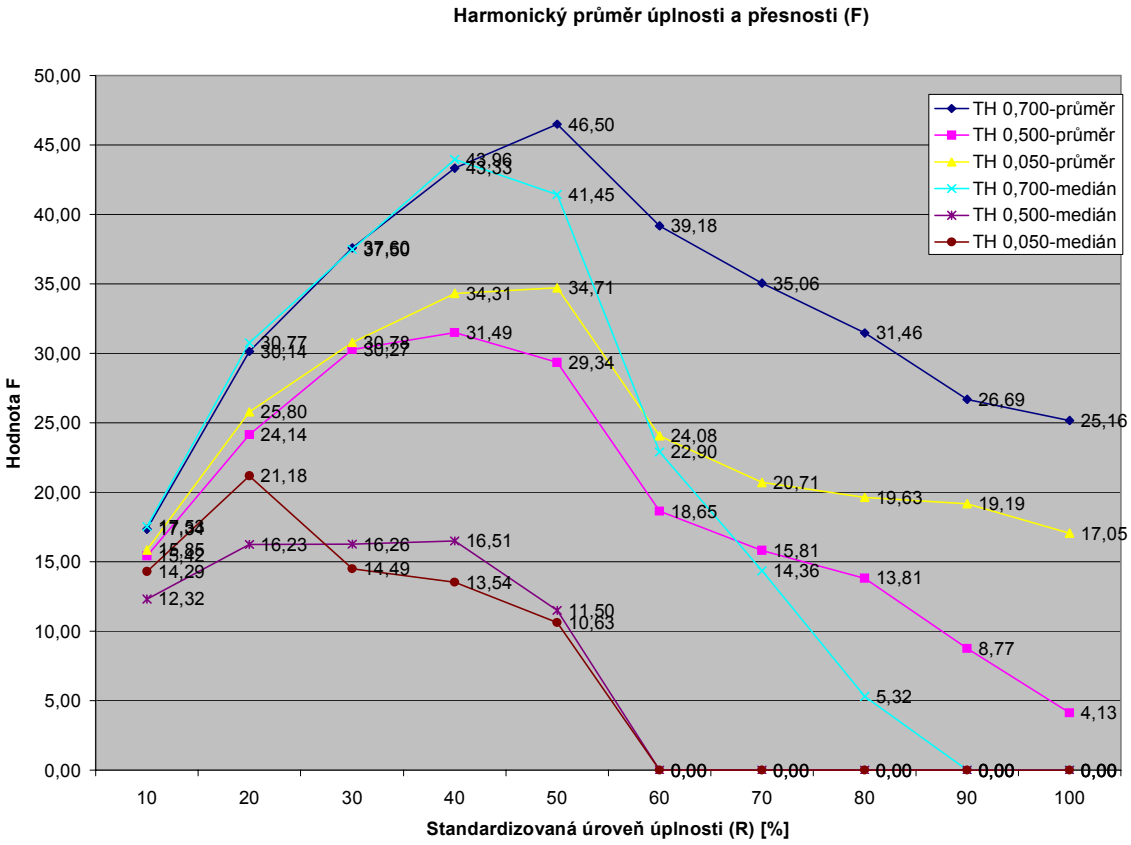
Tyto hypotézy by bylo třeba ověřit podrobnější analýzou testovacích dat a jednotlivých TMD, která v rámci tohoto testování nebyla na detailní úrovni provedena.

Harmonický průměr úplnosti a přesnosti

Tabulka č. 2 Harmonický průměr úplnosti a přesnosti

Standardní úroveň úplnosti [%]	*Harmonický průměr - průměr [%]*			*Harmonický průměr – medián [%]*		
	TH0700	*TH0500*	*TH0050*	*TH0700*	*TH0500*	*TH0050*
10	17,34	15,42	15,85	17,53	12,32	14,29
20	30,14	24,14	25,80	30,77	16,23	21,18
30	37,60	30,27	30,78	37,50	16,26	14,49
40	43,33	31,49	34,31	43,96	16,51	13,54
50	46,50	29,34	34,71	41,45	11,50	10,63
60	39,18	18,65	24,08	22,90	0,00	0,00
70	35,06	15,81	20,71	14,36	0,00	0,00
80	31,46	13,81	19,63	5,32	0,00	0,00
90	26,69	8,77	19,19	0,00	0,00	0,00
100	25,16	4,13	17,05	0,00	0,00	0,00

Graf č. 2



Výsledky této analýzy, založené na kombinaci úplnosti a přesnosti (hodnota míry F stoupá, pokud stoupají obě základní míry – úplnost a přesnost), potvrzují v případě parametru TH 0,700 předchozí zjištění: až do úrovně 50% úplnosti zvyšující se úplnost vyvažuje snižující se přesnost, za touto hranicí již snižující se přesnost redukuje výsledky získané zvyšující se úplností. Lze tedy říci, že efektivním způsobem lze vyhledat zhruba polovinu relevantních dokumentů, druhou polovinu relevantních dokumentů činí zejména snižující se přesnost obtížněji dostupnou.

Obdobné závěry platí i pro výsledky podle parametrů TH 0,500 a TH 0,050, pokud bereme v úvahu data vypočtená na základě průměrných hodnot.

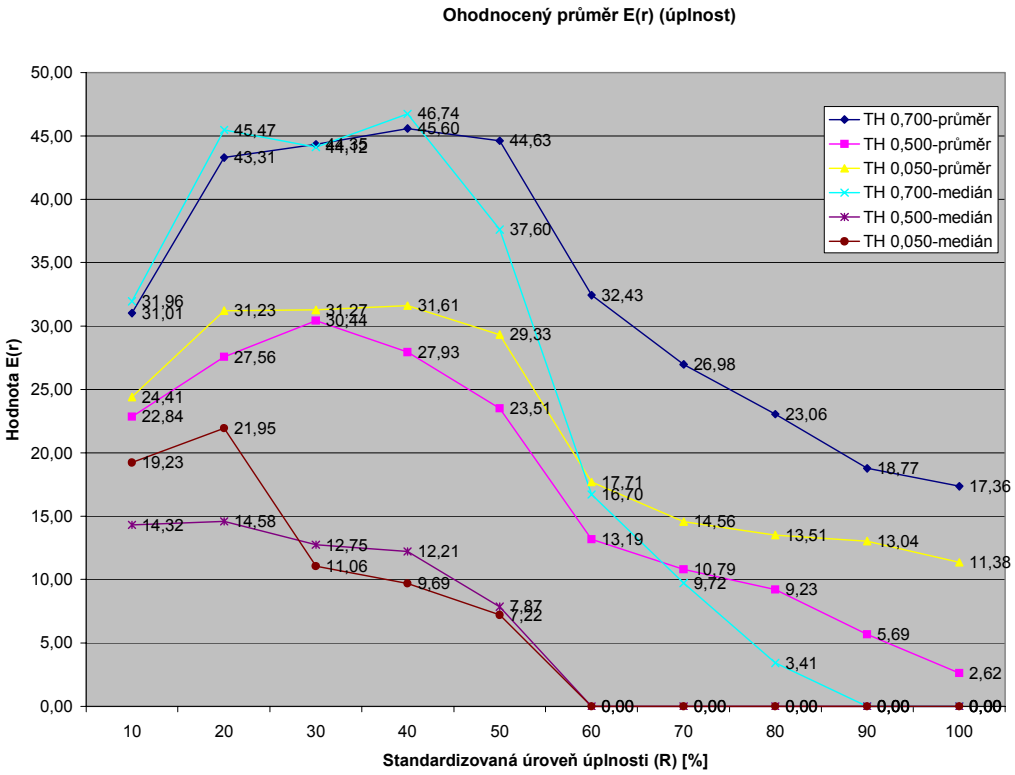
Ohodnocený průměr úplnosti a přesnosti

Obě tabulky i grafy již jen zdůrazňují zjištění podle předchozích výsledků. Uživatel preferující při vyhledávání úplnost, bude nejvíce uspokojen v rámci první poloviny výsledků vyhledávání, v druhé polovině výsledků je daní za zvyšující se úplnost podstatné snižování přesnosti. Uživatel preferující při vyhledávání přesnost bude nejvíce uspokojen při 50% úplnosti, za touto hranicí již jej mohou stále méně přesné výsledky přestat zajímat.

Tabulka č. 3 Harmonický průměr úplnosti a přesnosti - úplnost

Standardní úroveň úplnosti [%]	*Ohodnocený průměr/úplnost - průměr [%]*			*Ohodnocený průměr/úplnost – medián [%]*		
	TH0700	*TH0500*	*TH0050*	*TH0700*	*TH0500*	*TH0050*
10	31,01	22,84	24,41	31,96	14,32	19,23
20	43,31	27,56	31,23	45,47	14,58	21,95
30	44,35	30,44	31,27	44,12	12,75	11,06
40	45,60	27,93	31,61	46,74	12,21	9,69
50	44,63	23,51	29,33	37,60	7,87	7,22
60	32,43	13,19	17,71	16,70	0,00	0,00
70	26,98	10,79	14,56	9,72	0,00	0,00
80	23,06	9,23	13,51	3,41	0,00	0,00
90	18,77	5,69	13,04	0,00	0,00	0,00
100	17,36	2,62	11,38	0,00	0,00	0,00

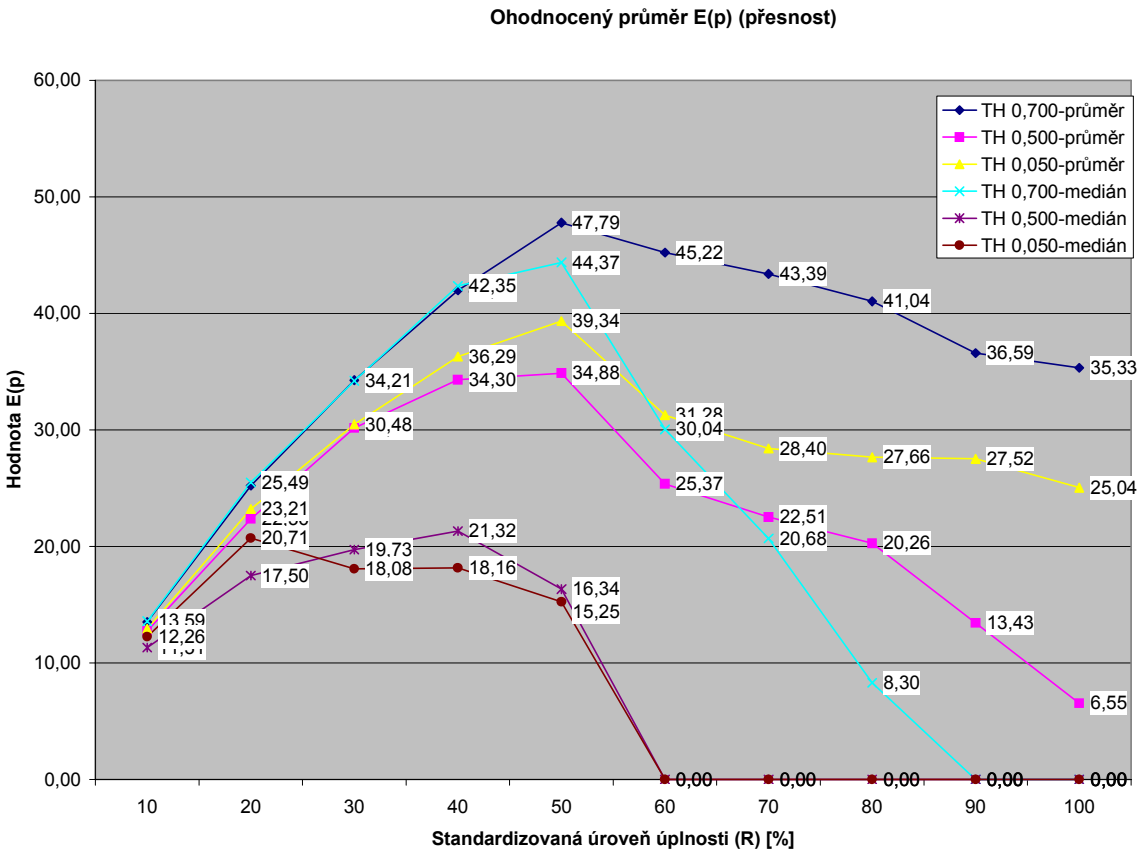
Graf č. 3



Tabulka č. 4 Harmonický průměr úplnosti a přesnosti - přesnost

Standardní úroveň úplnosti [%]	*Ohodnocený průměr/úplnost - průměr [%]*			*Ohodnocený průměr/úplnost – medián [%]*		
	TH0700	*TH0500*	*TH0050*	*TH0700*	*TH0500*	*TH0050*
10	13,52	12,76	12,94	13,59	11,31	12,26
20	25,22	22,36	23,21	25,49	17,50	20,71
30	34,26	30,17	30,48	34,21	19,73	18,08
40	41,98	34,30	36,29	42,35	21,32	18,16
50	47,79	34,88	39,34	44,37	16,34	15,25
60	45,22	25,37	31,28	30,04	0,00	0,00
70	43,39	22,51	28,40	20,68	0,00	0,00
80	41,04	20,26	27,66	8,30	0,00	0,00
90	36,59	13,43	27,52	0,00	0,00	0,00
100	35,33	6,55	25,04	0,00	0,00	0,00

Graf č. 4


Ohodnocený průměr E(p) (přesnost)
Hodnota E(p)
Standardizovaná úroveň úplnosti (R) [%]
TH 0,700-průměr
TH 0,500-průměr
TH 0,050-průměr
TH 0,700-medián
TH 0,500-medián
TH 0,050-medián
60,00
50,00
40,00
30,00
20,00
10,00
0,00
10
20
30
40
50
60
70
80
90
100
47,79
45,22
43,39
41,04
36,59
35,33
44,37
42,35
39,34
36,29
34,30
34,88
34,21
30,48
31,28
30,04
28,40
27,66
27,52
25,04
25,49
23,21
25,37
22,51
20,68
20,26
13,43
6,55
20,71
21,32
19,73
18,08
18,16
17,50
16,34
15,25
8,30
12,26
0,00
0,00
0,00
0,00
0,00

Reference

[1] BAEZA-YATES, R., RIBEIRO-NETO, B. Modern information retrieval. New York : ACM Press, 1999.

[2] BERRY, M. W., DUMAIS, S. T., LETSCHE., T. A. Computational Methods for Intelligent Information Access. Dostupný na URL: http://www.cs.utk.edu/~berry/sc95/sc95.html.

[3] BERRY, M.W., BROWNE. M. Understanding Search Engines. SIAM 1999.

[4] BINGHAM, E., MANNILA, H.. Random projection in dimensionality reduction: applications to image and text data. San Francisko : KDD, 2001.

[5] HÚSEK, D., SNÁŠEL, V. Neuronové sítě: přísliby a realita. *In* Databázy 96: 3. medzinárodný seminár o databázových technológiách. Bratislava: COFAX, 1996, s. 167-180

[6] PAPADIMITRIOU, C.H., RAGHAVAN, P., TAMAKI, H., VEMPALA, S. Latent Semantic Indexing: Probablistic Analysis. *Proc. 17th ACM Symp. on the Principles of Databases Systems*, 1998, s. 159-168.

[7] POKORNÝ, J., SNÁŠEL, V., HÚSEK, D. Dokumentografické informační systémy. Praha: Karolinum, 1998, s. 158.

[8] SALTON, G., MCGILL, M. Introduction to Modern Information Retrieval. New York: McGraw-Hill, 1983.

[9] SHUMSKY, S., YAROVOY, A. Assocative searching of textual information, Neuroinformatics - 99. Moscow : MIFI, 1999.

[10] SMETÁČEK, V. Sémantický analyzátor: základní pojmy a prvky. Olomouc: Universita Palackého, 1982.

[11] SMETÁČEK, V. Sémantický analyzátor: experimentální ověřování. Olomouc: Universita Palackého, 1984.

[12] SOSNA, K. Digitální knihovna Český parlament. In Infos 2000. Dostupný na URL: http://www.aib.sk/infos/infos2000/20.htm

[13] ÜBERLA, K. Faktorová analýza. Bratislava : Alfa, 1976.

[14] VELIČKOVÁ, H. Digitální knihovna Český parlament v roce 2000. Ikaros [online]. 2000, č. 10 [cit. 2000-10-01]. Dostupný na World Wide Web: <http://www.ikaros.cz/Clanek.asp?ID=200208216>.

[15] VRBÍKOVÁ, H. Jak je využívána elektronická knihovna Český parlament. Ikaros [online]. 1999, č. 05 [cit. 1999-05-01]. Dostupný na World Wide Web: <http://www.ikaros.cz/Clanek.asp?ID=200205068>.

[16] VRBÍKOVÁ, H. Projekt Elektronické knihovny v českém Parlamentu. Ikaros [online]. 1998, č. 07 [cit. 1998-07-01]. Dostupný na World Wide Web: <http://www.ikaros.cz/Clanek.asp?ID=200204017>.