Ukládání a vyhledávání textových informací

prof. RNDr. Jan Rauch, CSc.
Ing. Tomáš Kliegr, Ph.D.

Katedra informačního a znalostního inženýrství

Ukládání a vyhledávání textových informací

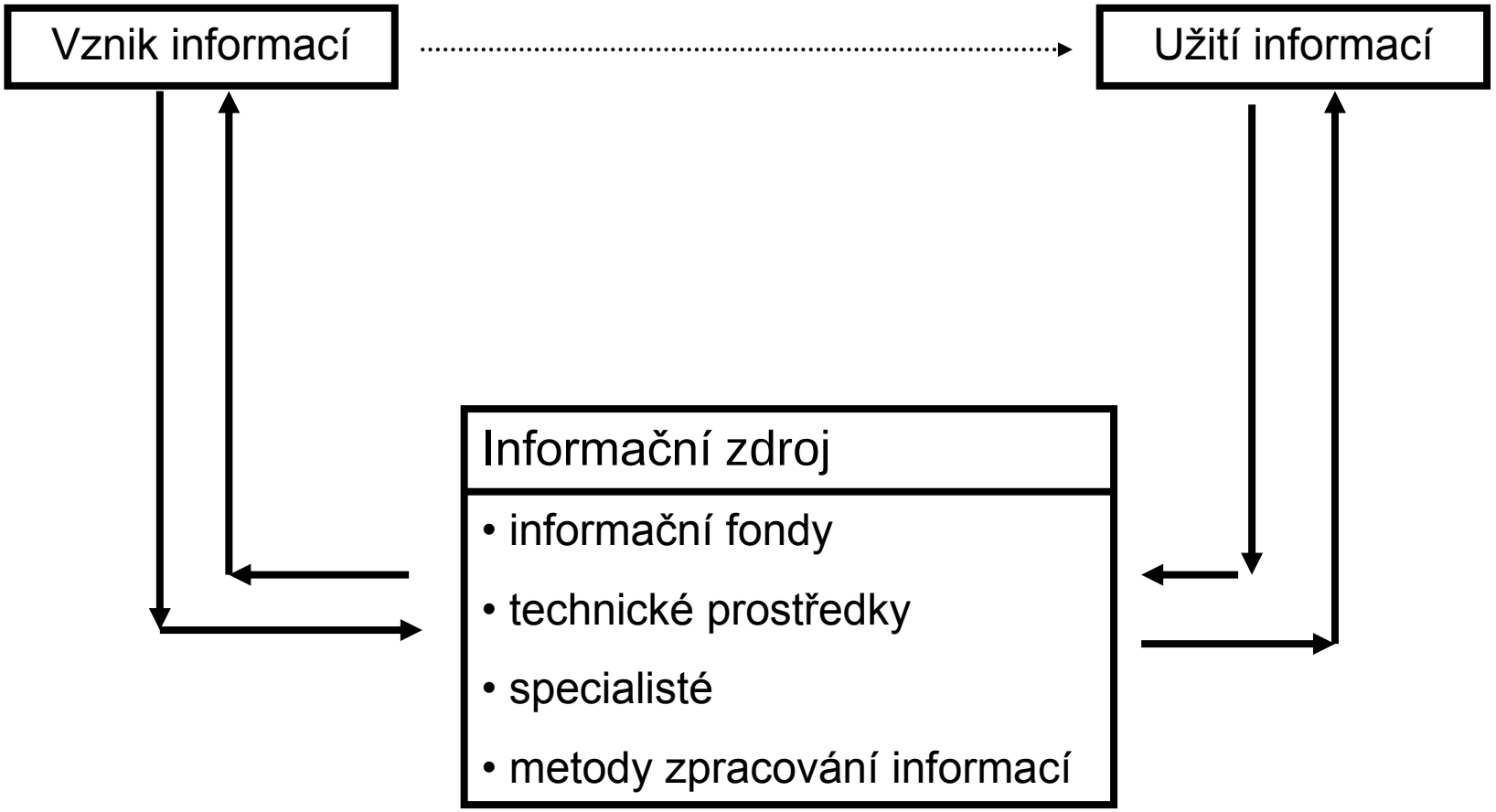
- Aplikace a souvislosti
  - Informační proces
  - Informační zdroj
  - Information retrieval a informační činnosti
- Témata
  - Vstupní zpracování
  - Hodnocení úspěšnosti ukládání a vyhledávání
  - Automatické indexování (automatická charakteristika obsahu)
  - Booleovský model a jeho rozšiřování
  - Invertovaný soubor
  - Vektorový model

Informační proces

VZNIK INFORMACÍ	**PŘEKÁŽKY**	UŽITÍ INFORMACÍ
obchod	čas	obchod
politika	prostor	politika
výzkum	struktura informací	výzkum
vývoj	věcná odbornost	vývoj
výroba	inform. odbornost	výroba
…	…	…

přenos informací: Informační zdroje

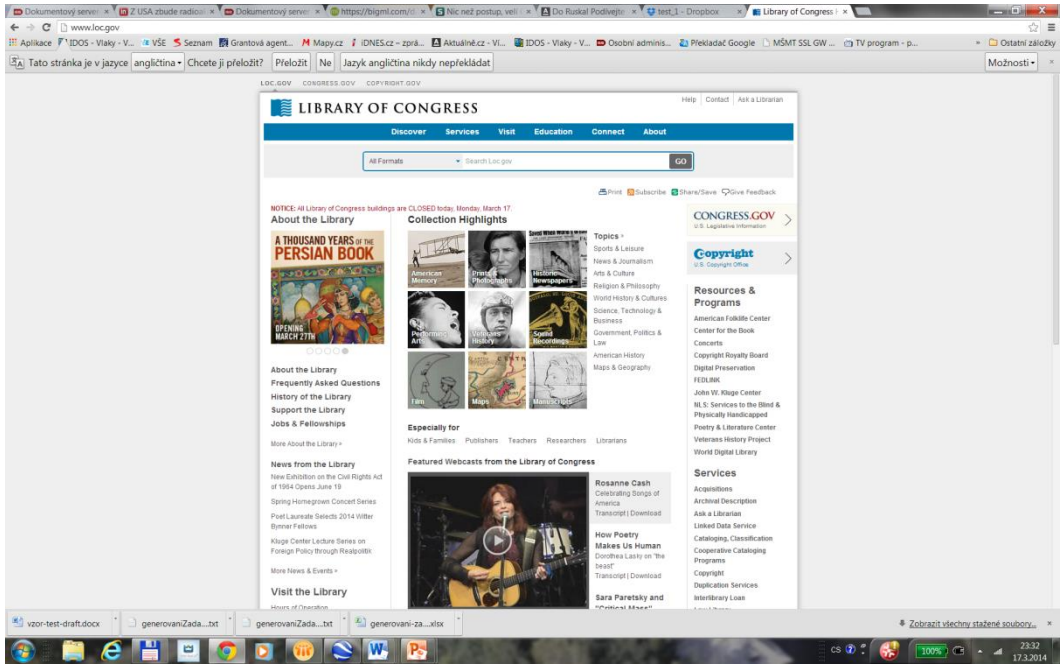
Informační zdroj



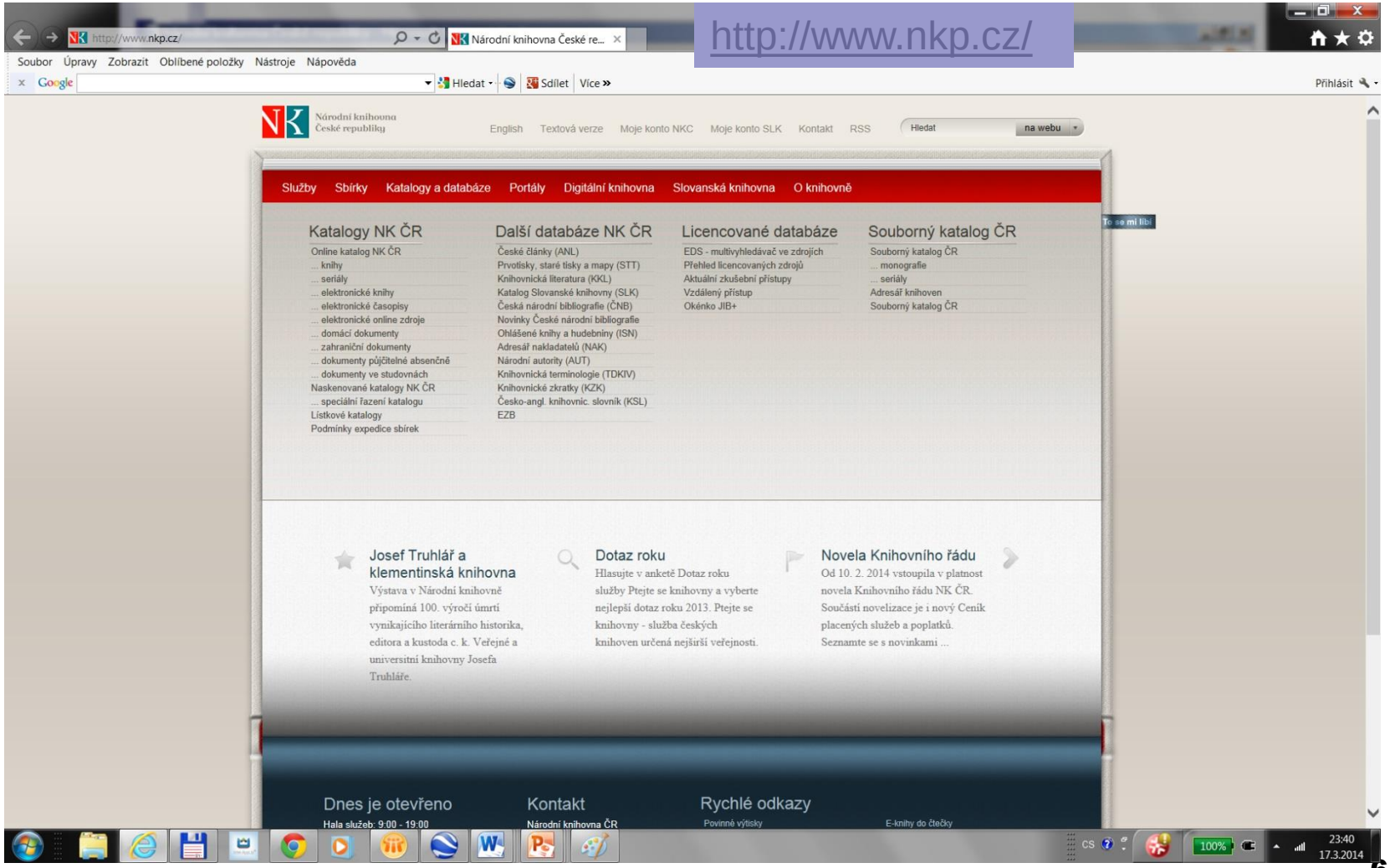
Knihovny, Specializované zdroje, Internet, …

The Library of Congress

- Oficiálně založena 1800, 900 knih tříděných podle formátu
- 1814 zničena, znovu vybudována s využitím soukromé knihovny Thomase Jeffersona
- 1992:
  - 90 milionů knihovních jednotek
  - 25 milionů knih
  - mapy, fotografie, rukopisy, magnetofonové pásky, …
  - http://www.loc.gov/index.html



Národní knihovna České republiky

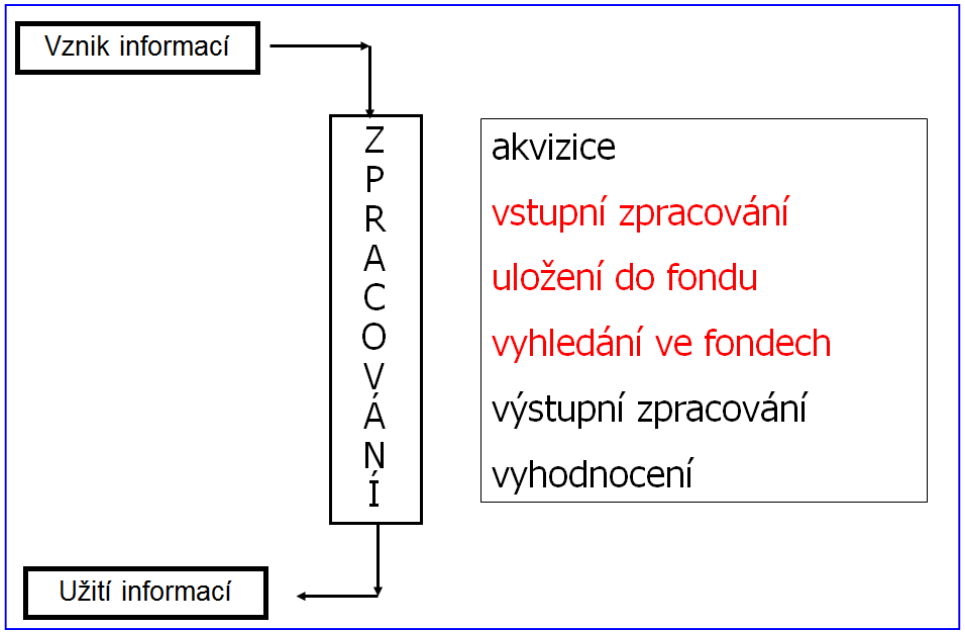


Information retrieval a informační činnosti

- Information retrieval (ukládání a vyhledávání informací):

  Pomoc uživatelům nalézt informace, které vyhovují jejich informačním potřebám

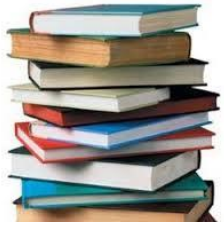
- Informační činnosti (zajímají nás zejména ty červené):

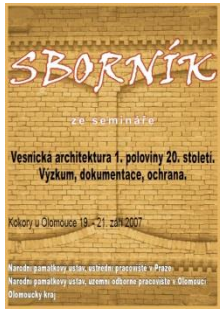


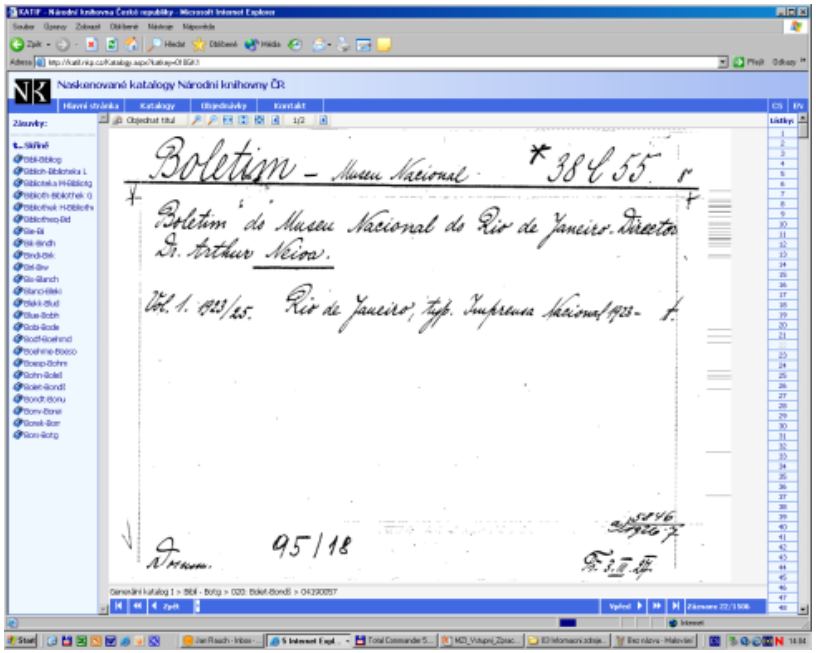
Témata

- Vstupní zpracování
- Hodnocení úspěšnosti ukládání a vyhledávání
- Automatické indexování (automatická charakteristika obsahu)
- Booleovský model a jeho rozšiřování
- Invertovaný soubor
- Vektorový model

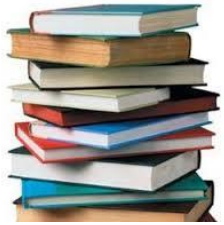
Vstupní zpracování – „ruční“ bibliografické záznamy

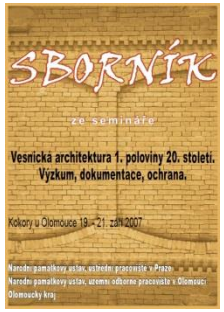




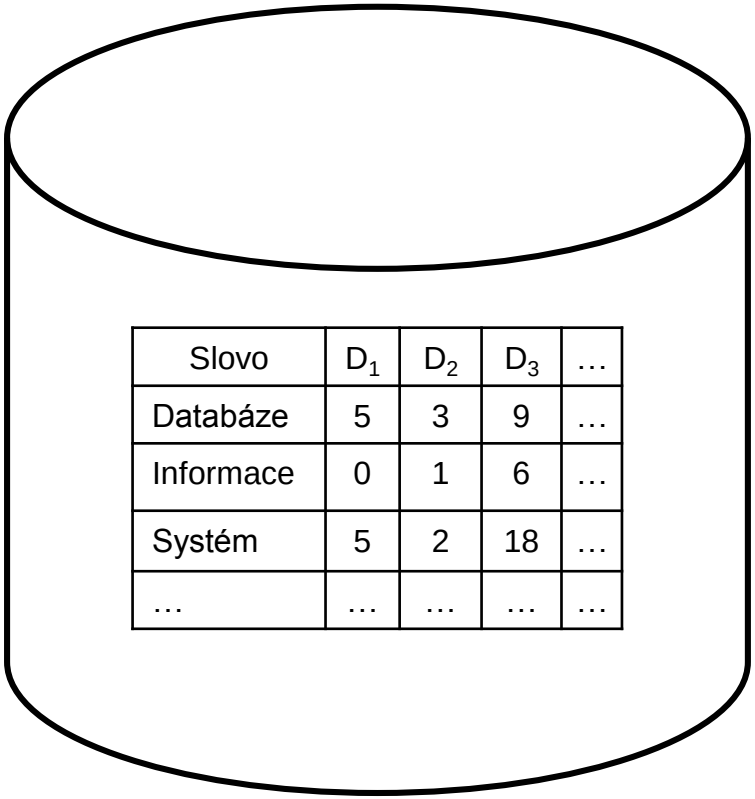


Vstupní zpracování – bibliografické záznamy



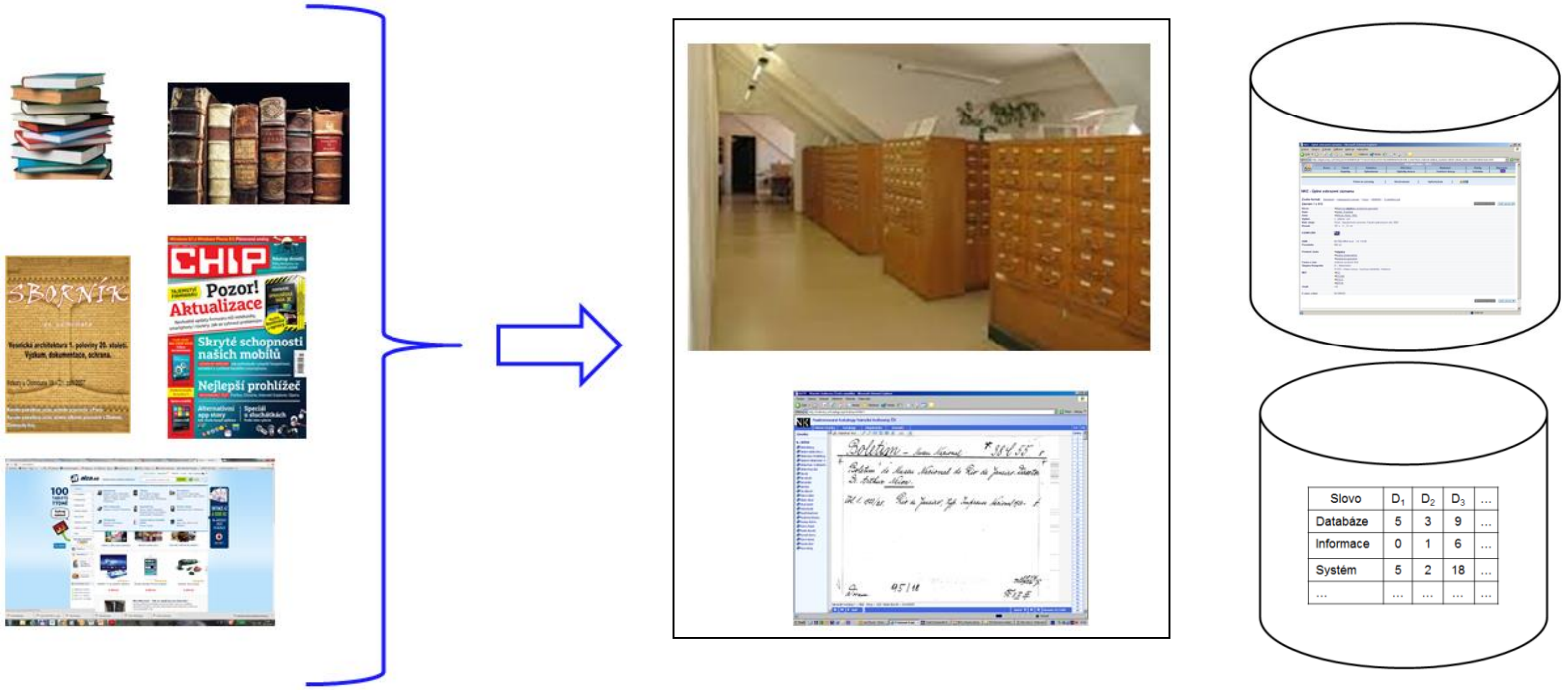


Vstupní zpracování – frekvence termínů

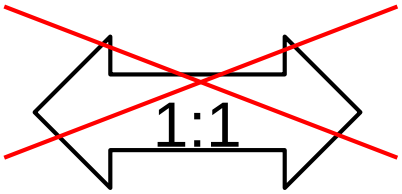


Slovo	D_1	D_2	D_3	…
Databáze	5	3	9	…
Informace	0	1	6	…
Systém	5	2	18	…
…	…	…	…	…

Vstupní zpracování – zdroj potíží



Dokument

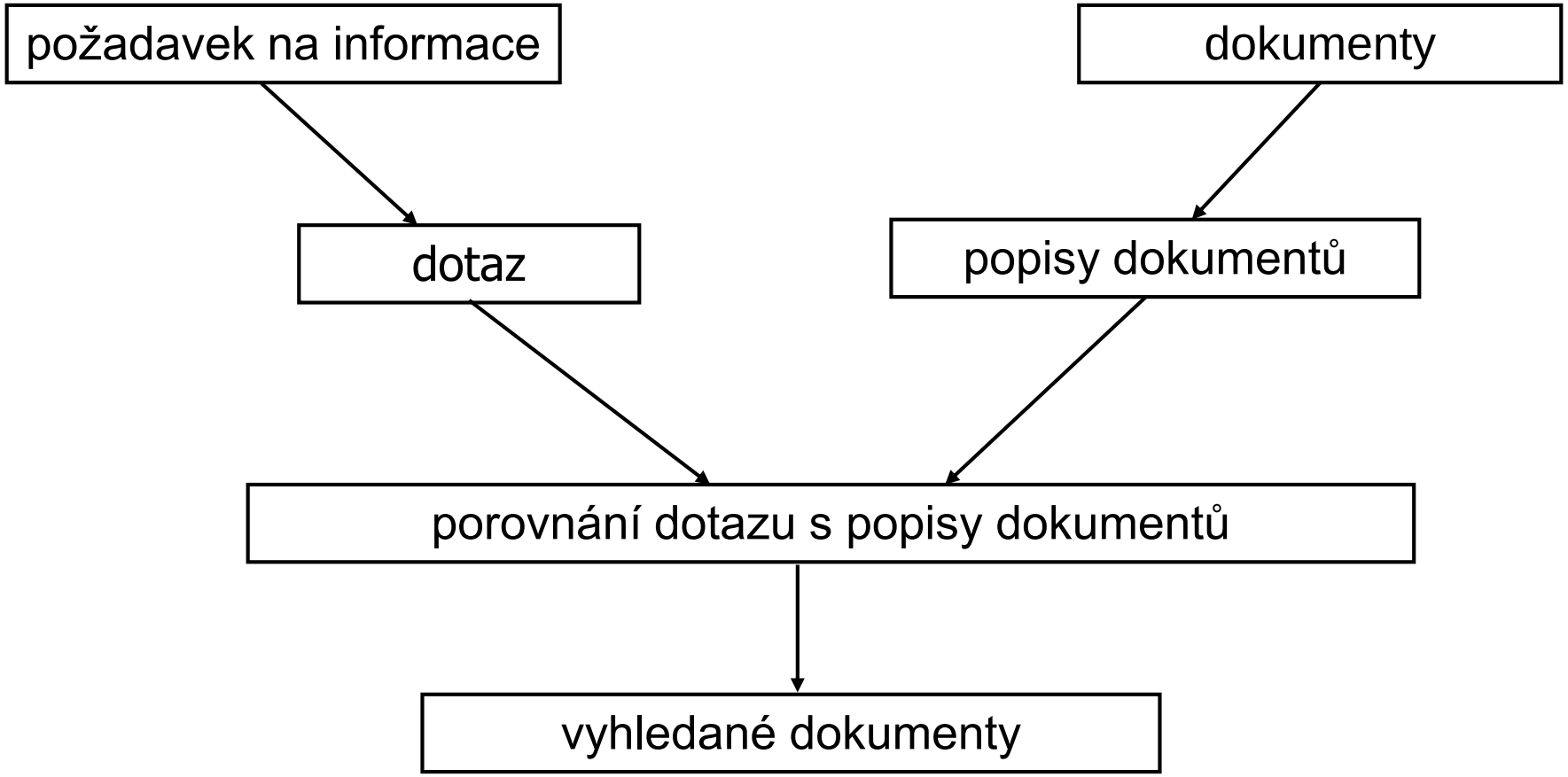


Popis dokumentu

Témata

- Vstupní zpracování
- Hodnocení úspěšnosti ukládání a vyhledávání
- Automatické indexování (automatická charakteristika obsahu)
- Booleovský model a jeho rozšiřování
- Invertovaný soubor
- Vektorový model

Ukládání a vyhledávání dokumentů



Ukládání a vyhledávání dokumentů – výsledky

DOKUMENTY	relevantní	irelevantní
vyhledané	a	b
nevyhledané	c	d

a = počet relevantních vyhledaných dokumentů

b = počet irelevantních vyhledaných dokumentů

c = počet relevantních nevyhledaných dokumentů

d = počet irelevantních nevyhledaných dokumentů

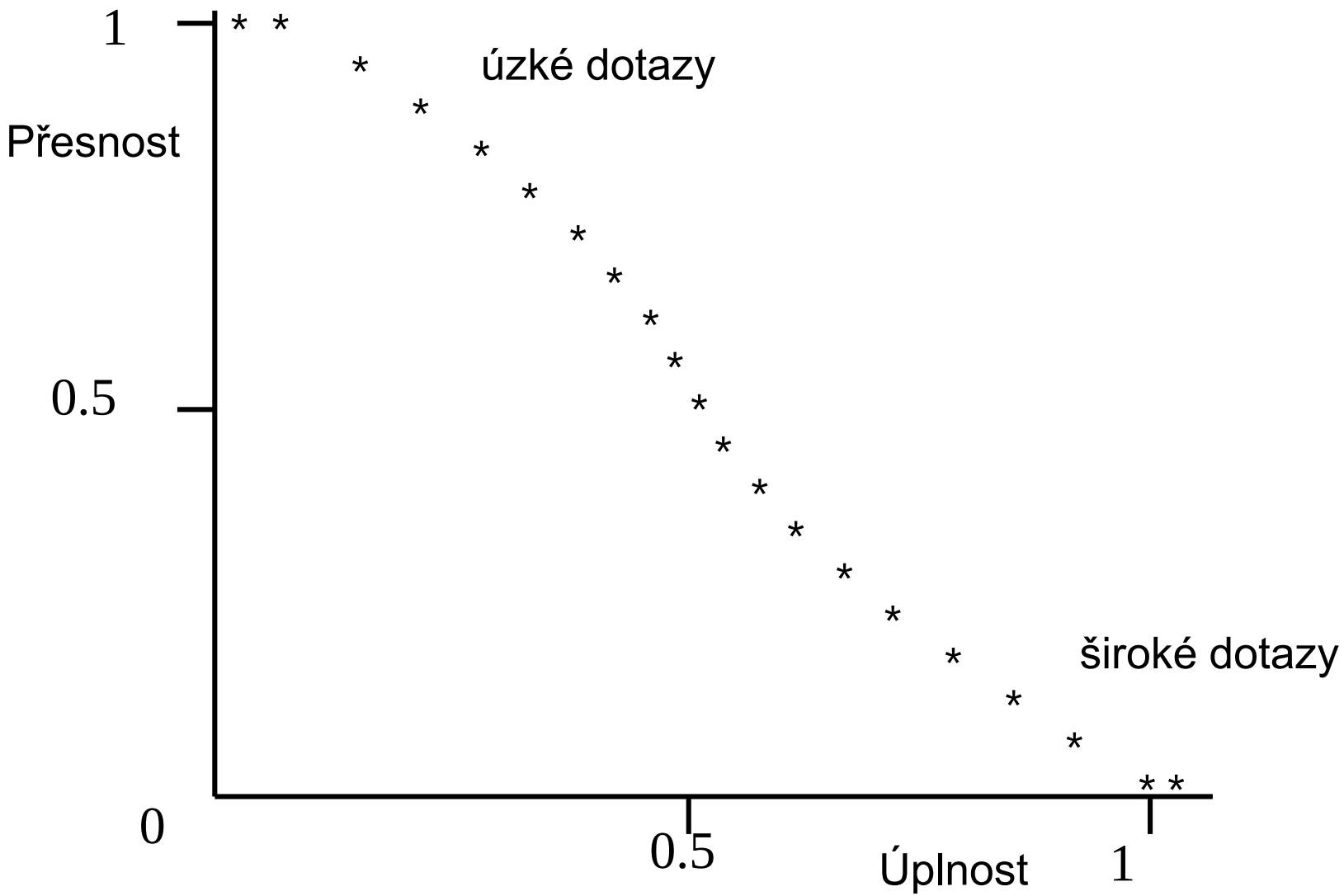
Úspěšnost vyhledávání dokumentů

DOKUMENTY	relevantní	irelevantní
vyhledané	a	b
nevyhledané	c	d

Přesnost (Precision) = $\frac{a}{a+b}$

Úplnost (Recall) = $\frac{a}{a+c}$

Vztah přesnosti a úplnosti



Témata

- Vstupní zpracování
- Hodnocení úspěšnosti ukládání a vyhledávání
- Automatické indexování (automatická charakteristika obsahu)
- Booleovský model a jeho rozšiřování
- Invertovaný soubor
- Vektorový model

Automatická charakteristika obsahu dokumentů

Literatura:

- Rauch, J.: Metody zpracování informací II, kapitola 5

- SALTON, G. - McGILL, M.: Introduction to Modern Information Retrieval. Tokyo, McGraw-Hill Book Company Japan 1983, 448 s.

Automatická charakteristika obsahu dokumentů

- důvody automatické charakteristiky obsahu
  - vyloučení subjektivity
  - rostoucí počet dokumentů
- jednoduchá indexovací metoda
- poznámky - problém specializovaného fondu
- modifikace jednoduché indexovací metody
- další přístupy

Jednoduchá indexovací metoda – příklad

Jednoduchá indexovací metoda – příklad

Stovky velryb uvázly na suchu. Proč?

Velryba - kandidát na klíčové slovo

Sydney, Praha - Stovka obyvatel Kingova ostrova ležícího mezi Austrálií a Tasmánií včera nešla do práce ani do školy. Den strávili na pláži, kde se snažili zvrátit osud téměř dvou stovek velryb a delfínů a alespoň některým z nich dát šanci na život. Celkem 194 mořských savců totiž uvázlo na pláži a nedokázalo se dostat zpět do moře. Na suchu postupně umírali. Ochranáři spolu s dobrovolníky přikrývali delfíny skákavé a kulohlavce černé dekami, jež pravidelně polévali vodou. Hloubili příkopy, kterými pak s pomocí vodních skútrů a motorových člunů mořské savce táhli do moře. Zdařilo se to u 48 z nich. Na 146 jich na pláži uhynulo.

V těchto končinách to není nic mimořádného. V listopadu zemřelo v Tasmánii 150 kulohlavců černých, když vypluli na pláž. Někdy dokonce odborníci přistupují k tomu, že savce sami uspí, aby je ušetřili pomalé smrti. Ochránci přírody tak o víkendu zabili vorvaňovce, který připlul až na pláž na severu novozélandského Aucklandu. Příliš rychlý odliv, nebo vina člověka?

Na vině může být prostě jen to, že zvířata u pevniny zastihl výraznější odliv a ona nestačila odplout, vědci však dávají stále častěji vinu lidem. Jejich zvukům, škodlivinám a v neposlední řadě navigaci. Velryby se pod vodou dorozumívají s pomocí zvuků. Ty jim slouží i pro orientaci. Kvůli stále hlasitějšímu působení člověka je pro ně však čím dál obtížnější se orientovat. Zvuk se navíc šíří mnohem hůře ve znečištěných vodách. I škodliviny vypouštěné do moří tak komplikují mořským savcům orientaci.

Říkáme tomu efekt koktejlového večírku," vysvětlil novozélandskému serveru nzherald. co britský vědec Mark Simmonds. Musíte mluvit hlasitěji a hlasitěji, aby vás bylo slyšet, až nakonec nikdo neslyší vůbec nic," popisuje.
A pak je tu známý problém sonarů. Ty totiž vydávají zvuky, které velrybám znějí povědomě. Některé mile, jiné výhrůžně. A podle toho se i chovají.

Sonar zní jako velryba na lovu a nutí savce k úprku „Vojenské sonary mohou vyvolat paniku u menších velryb a delfínů. Ti si totiž myslí, že jde o zpěv lovící kosatky, a prchají před ní," popisuje Christopher Clark z Cornellovy univerzity v New Yorku, který právě to, jak velryby vnímají jednotlivé zvuky, zkoumá už deset let. Úprk velryb pak podle Clarka může vést k tomu, že se savci dostanou až na mělčinu, z níž už nedokáží uniknout, a pomalu tam umírají. Vědec zjistil, že nízké frekvence lodních sonarů zní velrybám jako obr z říše velryb, keporkak. ……..

Jednoduchá indexovací metoda – příklad

Stovky velryb uvázly na suchu. Proč?

Nehodí se za klíčové slova – příklady

Sydney, Praha - Stovka obyvatel Kingova ostrova ležícího mezi Austrálií a Tasmánií včera nešla do práce ani do školy. Den strávili na pláži, kde se snažili zvrátit osud téměř dvou stovek velryb a delfínů a alespoň některým z nich dát šanci na život. Celkem 194 mořských savců totiž uvázlo na pláži a nedokázalo se dostat zpět do moře. Na suchu postupně umírali. Ochranáři spolu s dobrovolníky přikrývali delfíny skákavé a kulohlavce černé dekami, jež pravidelně polévali vodou. Hloubili příkopy, kterými pak s pomocí vodních skútrů a motorových člunů mořské savce táhli do moře. Zdařilo se to u 48 z nich. Na 146 jich na pláži uhynulo.

V těchto končinách to není nic mimořádného. V listopadu zemřelo v Tasmánii 150 kulohlavců černých, když vypluli na pláž. Někdy dokonce odborníci přistupují k tomu, že savce sami uspí, aby je ušetřili pomalé smrti. Ochránci přírody tak o víkendu zabili vorvaňovce, který připlul až na pláž na severu novozélandského Aucklandu. Příliš rychlý odliv, nebo vina člověka?

Na vině může být prostě jen to, že zvířata u pevniny zastihl výraznější odliv a ona nestačila odplout, vědci však dávají stále častěji vinu lidem. Jejich zvukům, škodlivinám a v neposlední řadě navigaci. Velryby se pod vodou dorozumívají s pomocí zvuků. Ty jim slouží i pro orientaci. Kvůli stále hlasitějšímu působení člověka je pro ně však čím dál obtížnější se orientovat. Zvuk se navíc šíří mnohem hůře ve znečištěných vodách. I škodliviny vypouštěné do moří tak komplikují mořským savcům orientaci.

Říkáme tomu efekt koktejlového večírku," vysvětlil novozélandskému serveru nzherald s. Musíte mluvit hlasitěji a hlasitěji, aby vás bylo slyšet, až nakonec nikdo neslyší vůbec A pak je tu známý problém sonarů. Ty totiž vydávají zvuky, které velrybám znějí pově A podle toho se i chovají.

Sonar zní jako velryba na lovu a nutí savce k úprku „Vojenské sonary mohou vyvolat p Ti si totiž myslí, že jde o zpěv lovící kosatky, a prchají před ní," popisuje Christopher Cla Yorku, který právě to, jak velryby vnímají jednotlivé zvuky, zkoumá už deset let. Úprk v k tomu, že se savci dostanou až na mělčinu, z níž už nedokáží uniknout, a pomalu tam umírají. Vědec zjistil, že nízké frekvence lodních sonarů zní velrybám jako obr z říše velryb, keporkak. ……..

Jednoduchá indexovací metoda

Princip:

Jestliže se slovo vyskytuje v dokumentu s dostatečnou frekvencí, pak se dokument týká pojmu odpovídajícímu tomuto slovu.

Vstup:

N dokumentů $D_1,\ldots, D_N$

Výstup:

Klíčová slova pro každý dokument

Jednoduchá indexovací metoda - postup

1) Vynechej stop slova.

2) Spočti frekvence zbývajících slov S_1, …, S_K.

$F_{i,j}$ – frekvence slova S_j v dokumentu D_i

3) Zvol prahovou hodnotu P.

S_j je klíčové slovo pro D_i právě když $F_{i,j} > P$

Jednoduchá indexovací metoda - příklad

Dokumenty:

D_1 Novák: Vyhledávání informací pomocí počítačů.

D_2 Kadlec: Historie počítačů.

...

$D_{10\,000}$ Kovář: Informace o firmách

prahová hodnota P = 6

	S_1	S_2	S_3	S_4	S_5	S_6	…
	počítač	informace	vyhledávání	historie	systém	firma	…
D_1	12	15	9	1	5	0	…
D_2	11	4	1	13	5	1	…
…	…	…	…	…	…	…	…
$D_{10\,000}$	5	13	8	1	3	15	…

Jednoduchá indexovací metoda - poznámky

- Častý výskyt slova znamená, že dokument se týká tématu odpovídajícího tomuto slovu.
- Jestliže se dokument týká tématu odpovídajícího nějakému slovu, pak se toto slovo v dokumentu vyskytuje s velkou frekvencí.
- Slova s vysokou frekvencí nemusí rozlišit dokumenty na relevantní a irelevantní. („Počítač" ve fondu informatické literatury).

Modifikace jednoduché indexovací metody

Cíl: Klíčová slova

- charakterizující obsah
- oddělující dokumenty

Princip:

- vyjádříme stupeň kterým slovo S_j charakterizuje obsah dokumentu D_i :

  $F_{i,j}$ - frekvence slova S_j v dokumentu D_i

- vyjádříme stupeň kterým slovo S_j odděluje dokumenty: $\log (N/DF_j)$

  DF_j – počet dokumentů obsahujících S_j

- použijeme váhu $W_{i,j} = F_{i,j} * \log (N/DF_j)$ místo frekvence $F_{i,j}$

Modifikovaná jednoduchá indexovací metoda – postup

1) Vynechej stop slova.

2) Spočti váhy $W_{i,j}$ zbývajících slov S_1, …, S_K.

$W_{i,j}$ – váha slova S_j v dokumentu D_i

3) Zvol prahovou hodnotu P.

S_j je klíčové slovo pro D_i právě když $W_{i,j} > P$

Váhy slov – příklad

N = 10 000, DF_j = počet dokumentů se slovem S_j

$F_{i,j}$ = frekvence slova S_j pro dokument D_i , $W_{i,j} = F_{i,j} * \log(N/DF_j)$, práh P = 6

Slovo S_j	DF_j	%	N/DF_j	$\log(N/DF_j)$	$F_{i,j}$	$W_{i,j}$
Databáze	10	0.1	1 000	3	1	3
					2	6
					5	15
Uživatel	30	0.3	333.3	2.52	1	2.5
					3	7.6
					5	12.6
Tiskárna	100	1.0	100	2	1	2
					3	6
					5	10
Metoda	500	5.0	20	1.3	1	1.3
					5	6.5
Počítač	2000	20	5	0.7	1	0.7
					9	6.3
Informace	9000	90	1.1	0.15	1	0,1
					42	6,1

Témata

- Vstupní zpracování
- Hodnocení úspěšnosti ukládání a vyhledávání
- Automatické indexování (automatická charakteristika obsahu)
- Booleovský model a jeho rozšiřování
- Invertovaný soubor
- Vektorový model

Booleovský model a jeho rozšiřování

Literatura:

Rauch, J.: Metody zpracování informací II, kapitoly 6, 8

SALTON, G. - McGILL, M.: Introduction to Modern Information Retrieval. Tokyo, McGraw-Hill Book Company Japan 1983, 448 s.

Booleovský model a jeho rozšiřování

- Booleovská logika
- Základní vlastnosti booleovského modelu – přehled
- Důvody rozšiřování booleovského modelu - experiment se Stairs
- Cíle rozšiřování
- Rozšíření pomocí fuzzy logiky
- Geometrické rozšíření

Booleovská logika - 1

- Booleovská (výroková) logika se zabývá výroky a jejich pravdivostí.
- Každý výrok je buď pravdivý nebo nepravdivý
- Rozlišujeme základní a složené výroky
- Pravdivost základních výroků je dána vnějšími okolnostmi
- Složené výroky se vytvářejí pomocí výrokových spojek
- Používají se pravdivostní tabulky pro výrokové spojky

Booleovská logika - 2

Příklady složených výroků:

$\neg U, \quad U \wedge V, \quad U \vee V, \quad U \wedge (V \vee W)$

U, V, W jsou základní výroky

U	V	$U \wedge V$	$U \vee V$	$\neg U$
1	1	1	1	0
1	0	0	1	0
0	1	0	1	1
0	0	0	0	1

Booleovská logika - 3

vyhodnocení složeného výroku $U \wedge (V \vee W)$

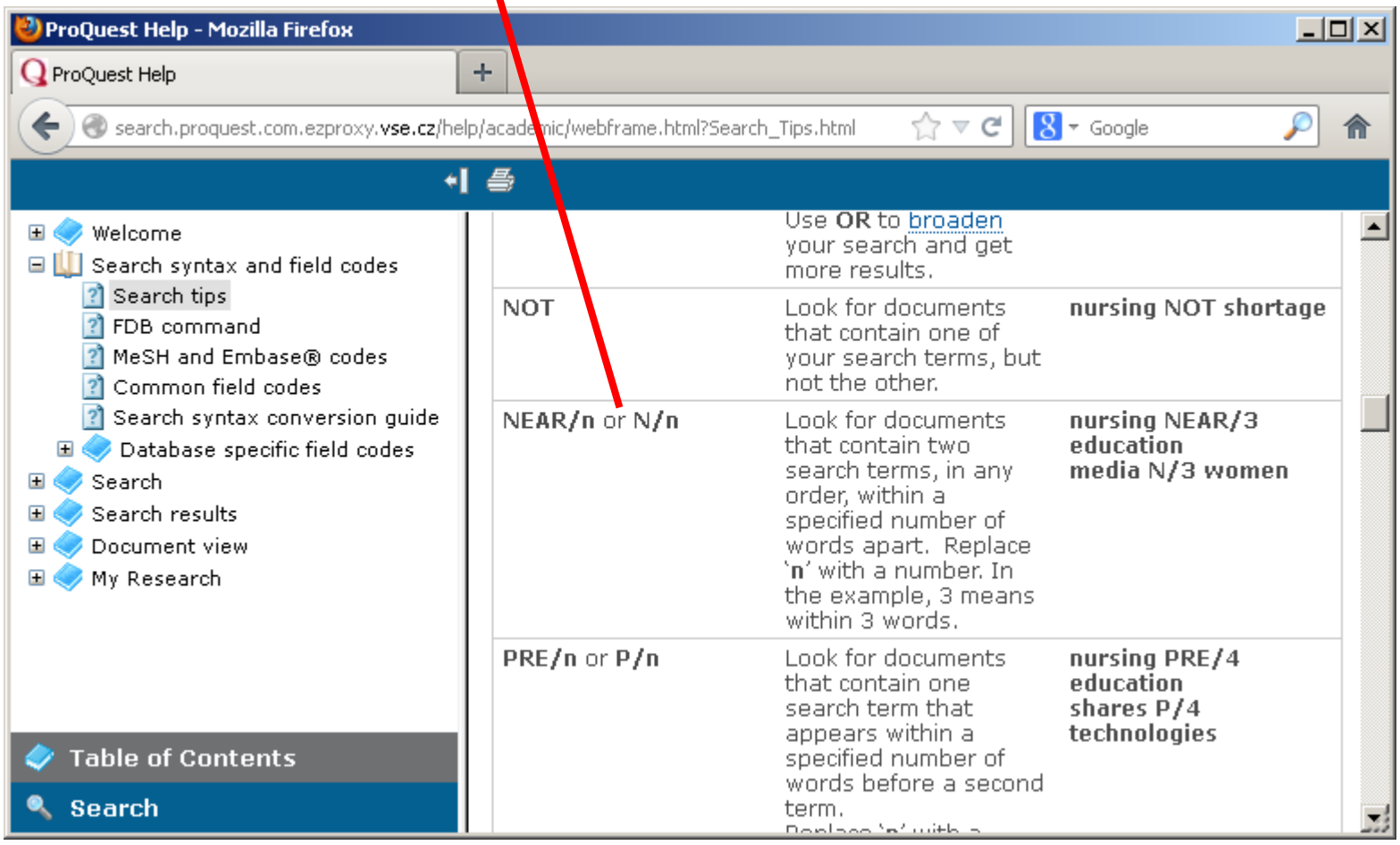
U	V	W	$V \vee W$	$U \wedge (V \vee W)$
1	1	1	1	1
1	1	0	1	1
1	0	1	1	1
1	0	0	0	0
0	1	1	1	0
0	1	0	1	0
0	0	1	1	0
0	0	0	0	0

Booleovský model vyhledávání dokumentů – přehled

- Pro bibliografické záznamy i záznamy (s úseky) plných textů
- Základní výroky se týkají výskytu výrazů
- Používají se logické spojky AND, OR, NOT
- A NOT B znamená A AND NOT B
- Jsou k dispozici vzdálenostní (proximitní) operátory
- Je k dispozici pravostranné rozšíření
- Různé systémy mají různé další možnosti

Booleovský model vyhledávání dokumentů – příklad

ProQuest



Experiment se STAIRS (1985)

40 000 právnických textů

- soudní případy, protokoly, výslechy
- celkem 350 000 stran

51 požadavků – podklady pro případy

přání: úplnost ~ 75%

výsledek: přesnost ~ 80%

úplnost ~ 20% !

Cíle rozšiřování Booleovského modelu

- Rozlišení důležitosti deskriptorů v dokumentu
- Rozlišení důležitosti deskriptorů v dotazu
- Řazení vybraných dokumentů podle důležitosti
- Odstranění tvrdosti booleovských operací (AND)

Booleovský model a jeho rozšiřování

- Booleovská logika
- Základní vlastnosti booleovského modelu – přehled
- Přesnost a úplnost v booleovském modelu
- Důvody rozšiřování booleovského modelu - experiment se Stairs
- Cíle rozšiřování
- Rozšíření pomocí fuzzy logiky
- Geometrické rozšíření

Fuzzy logika

- Fuzzy = chomáčovitý, chmýřivý, kučeravý, zakalený, nalíznutý, matný, mlhavý, neostrý, …
- Fuzzy logika je konzervativním rozšířením booleovské logiky
- Připouští různé úrovně pravdivosti
- Pr ("Míč je veliký") = 0.6
- Pr („CD-ROM je deskriptorem pro dokument A") = 0.9

Pravdivost složených výroků ve fuzzy logice

Pr(U)	Pr(V)	Pr (U OR V) max(Pr(U), Pr(V))	Pr (U AND V) min(Pr(U), Pr(V))	Pr(NON U) 1 - Pr(U)
1	1	1	1	0
1	0.7	1	0.7	0
0.7	0.4	0.7	0.4	0.3
0.7	0	0.7	0	0.3
1	0	1	0	0
0	1	1	0	1
0	0.4	0.4	0	1
0	0	0	0	1

Váha deskriptoru v dokumentu a v dotazu - příklady

dokument	váha deskriptorů U	váha deskriptorů V	výsledná váha dotazu [U; 0.7] OR [V; 0.9]	výsledná váha dotazu [U; 0.7] AND [V; 0.9]
D_1	1	1	0.9	0.7
D_2	1	0	0.7	0.0
D_3	0.6	0.8	0.72	0.42
D_4	0	0.9	0.81	0.0

Výpočet pro D_3 :

Váha ([U; 0.7] OR [V; 0.9]) = max (0.6 * 0.7, 0.8 * 0.9) = 0.72

Váha ([U; 0.7] AND [V; 0.9]) = min(0.6 * 0.7, 0.8 * 0.9) = 0.42

Složené výroky ve fuzzy logice – příklady

Pr(U)	Pr(V)	Pr (W)	Pr(U OR V OR W)	Pr (U AND V AND W)
1	1	1	1	1
0.1	0.9	0.2	0.9	0.1
0.8	0.9	0.8	0.9	0.8
0	0.9	0.1	0.9	0
0.9	0.9	0.1	0.9	0.1
0.7	0.3	0.1	0.7	0.1
0.1	0.1	0.1	0.1	0.1

Rozšíření booleovského modelu pomocí fuzzy logiky

- Rozlišení důležitosti deskriptorů v dokumentu – vyřešeno
- Rozlišení důležitosti deskriptorů v dotazu – vyřešeno
- Řazení vybraných dokumentů podle důležitosti – vyřešeno
- Odstranění tvrdosti booleovských operací (AND) – NE!

Následuje kvízová otázka

Kvíz – Otázka 1

	váha deskriptorů		výsledná váha dotazu	
Dokument	U	V	[U; 0.5] OR [V; 0.5]	[U; 0.5] AND [V; 0.5]
D	1.0	0.1	u	v

Které tvrzení je pravdivé?

u = 0.5 a v = 0.5

u = 0.5 a v = 0.05

u = 0.05 a v = 0.5

u = 0.05 a v = 0.05

u = 0.5 a v = 0.0

u = 0.0 a v = 0.5

u = 0.5 a v = 0.1

u = 0.1 a v = 0.5

Booleovský model a jeho rozšiřování

- Booleovská logika
- Základní vlastnosti booleovského modelu – přehled
- Přesnost a úplnost v booleovském modelu
- Důvody rozšiřování booleovského modelu - experiment se Stairs
- Cíle rozšiřování
- Rozšíření pomocí fuzzy logiky
- Geometrické rozšíření

Geometrické rozšíření booleovského modelu

Cíl: Odstranit tvrdost booleovských operací

Princip:

- Připouští váhy slov v dotazu i v dokumentu
- Dokument = bod v prostoru
- Hodnota (U AND V) $\leq$ Hodnota (U OR V)

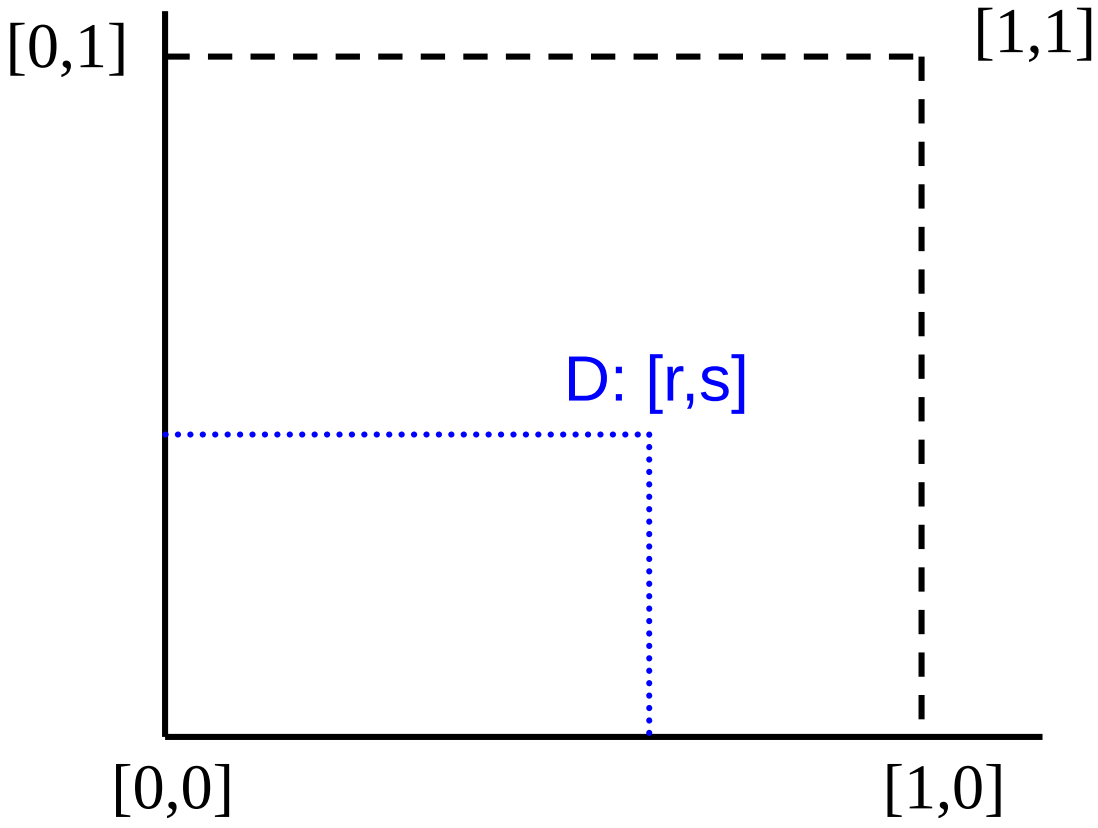
Příklad:

Dvě klíčová slova: $\rightarrow$ prostor = rovina

Dokument – bod v rovině

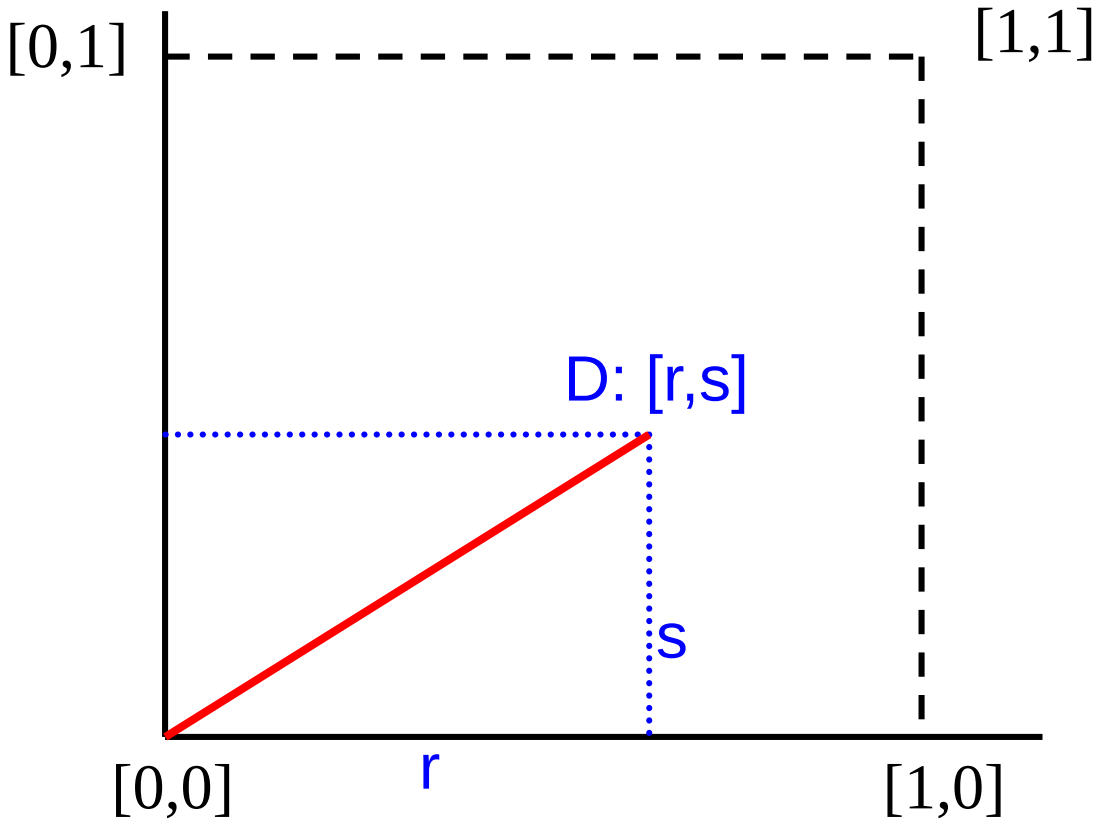
Dokument D - klíčová slova: **U** s vahou **r**

V s vahou **s**



Hodnota (U OR V)

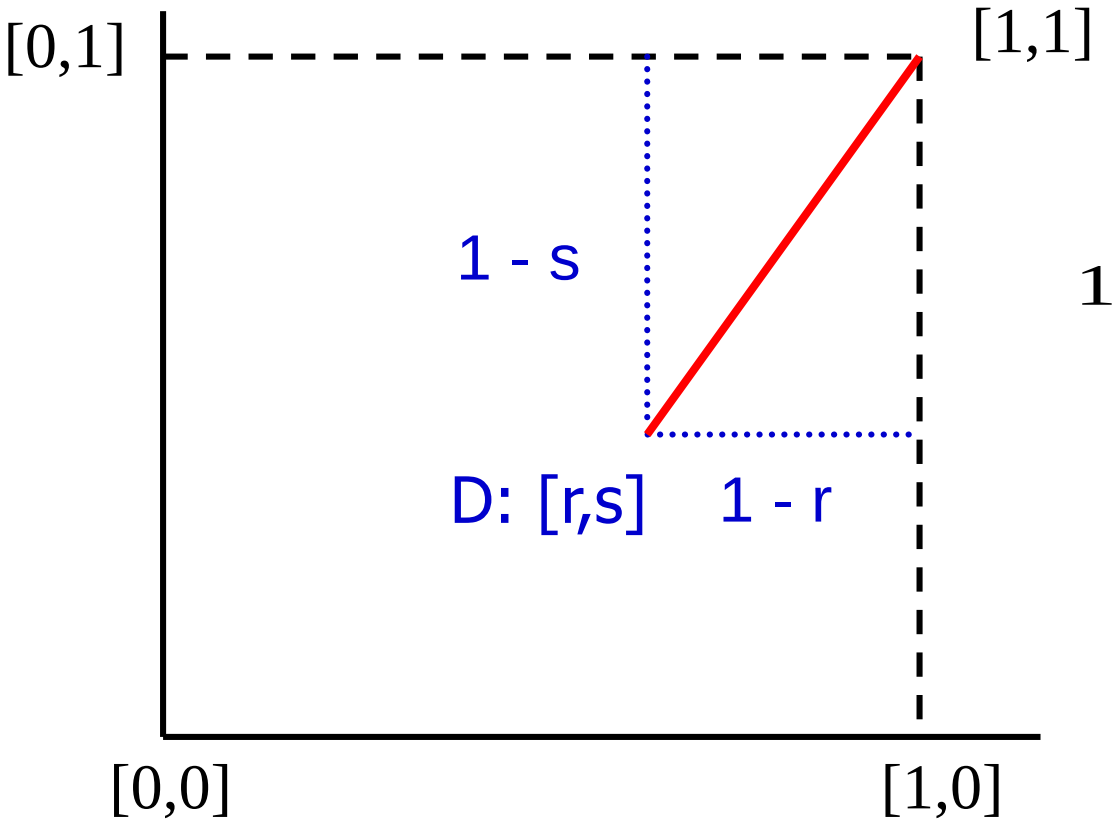
Hodnota (U OR V) - přímo úměrná vzdálenosti [r,s] od [0,0]



$$\sqrt{\frac{r^2 + s^2}{2}}$$

Hodnota (U AND V)

Hodnota (U AND V) - nepřímo úměrná vzdálenosti [r,s] od [1,1]



$$1 - \sqrt{\frac{(1-r)^2 + (1-s)^2}{2}}$$

Složené výroky v geometrickém rozšíření – příklad

Dokument D,

deskriptory U a V s váhou v dokumentu

váha U	váha V	Hodnota (U OR V)	hodnota (U AND V)
1	1	1	1
1	0	0.7	0.3
0.3	0.8	0.6	0.5
0	1	0.7	0.3
0	0	0	0

Geometrické rozšíření booleovského modelu

- Rozlišení důležitosti deskriptorů v dokumentu – vyřešeno
- Rozlišení důležitosti deskriptorů v dotazu – vyřešeno
- Řazení vybraných dokumentů podle důležitosti – vyřešeno
- Odstranění tvrdosti booleovských operací (AND) – vyřešeno

Porovnání standardního modelu s rozšířenými

			Přesnost pro konstantní úplnost		
Fond	dokumentů	dotazů	Booleovský	Fuzzy	Geom. rozš.
CACM	3 024	52	0.179	0.156	0.331
				-14%	+72%
CISI	1 460	35	0.119	0.100	0.180
				-11%	+62%
INSPEC	12 684	77	0.116	0.131	0.270
				+13%	+133%
MED	1 033	30	0.207	0.237	0.557
				+15%	+167%

Oprava

Ve skriptech :

Rauch, J.: Metody zpracování informací II,

Odstavec 8.4, str. 49 má být

hodnota ([U,a] AND [V,b]) = $1-\sqrt{\frac{a^2*(1-r)^2+b^2*(1-s)^2}{a^2+b^2}}$

místo

hodnota ([U,a] AND [V,b]) = $\sqrt{\frac{a^2*(1-r)^2+b^2*(1-s)^2}{a^2+b^2}}$

Témata

- Vstupní zpracování
- Hodnocení úspěšnosti ukládání a vyhledávání
- Automatické indexování (automatická charakteristika obsahu)
- Booleovský model a jeho rozšiřování
- Invertovaný soubor
- Vektorový model

Datové struktury a algoritmy

Literatura:

Rauch, J.: Metody zpracování informací II, kapitola 14

oprava: str. 83, má být $\{2\} \cup \{1,3\} = \{1,2,3\}$ místo $2 \cup 1,3 = 1,2,3$

Invertovaný soubor – princip

Záznamy dokumentů

Novák	1	Kadlec	2	Zouhar	3	Květnatá	4
Počítače a informace		Metody vyhledávání		Operační systémy		Ukládání informací	
GRADA 2010		PASEKA 2011		UK 2013		VŠE 2008	
počítač informace vyhledávání		informace vyhledávání metoda		počítač systém tiskárna		informace systém ukládání	

Slova v invertovaném souboru

- podle abecedy
- u každého slova seznam dokumentů

informace	1,2,4
metoda	2
počítač	1,3
systém	3,4
tiskárna	3
ukládání	4
vyhledávání	1,2

Zde pro jednoduchost použita pouze klíčová slova

Invertovaný soubor – Booleovské dotazy

Slova v invertovaném souboru
- podle abecedy
- u každého slova seznam dokumentů

informace	1,2,4
metoda	2
počítač	1,3
systém	3,4
tiskárna	3
ukládání	4
vyhledávání	1,2

informace AND metoda : $\{1,2,4\} \cap \{2\} = \{2\}$

metoda OR počítač : $\{2\} \cup \{1,3\} = \{1,2,3\}$

informace AND NOT ukládání : $\{1,2,4\} - \{4\} = \{1,2\}$

Invertovaný soubor – pravostranné rozšíření

Příklad dotazu: inform* AND metoda

Úsek invertovaného souboru:

…	…
ikona	2, 8, 9
imaginární	1,3, 6, 8
informace	3,4
informatika	3
informatizace	4
ironie	1,2
…	…

Rozepsaný dotaz:

(informace AND metoda) OR (informatika AND metoda) OR (informatizace AND metoda)

Invertovaný soubor – levostranné rozšíření

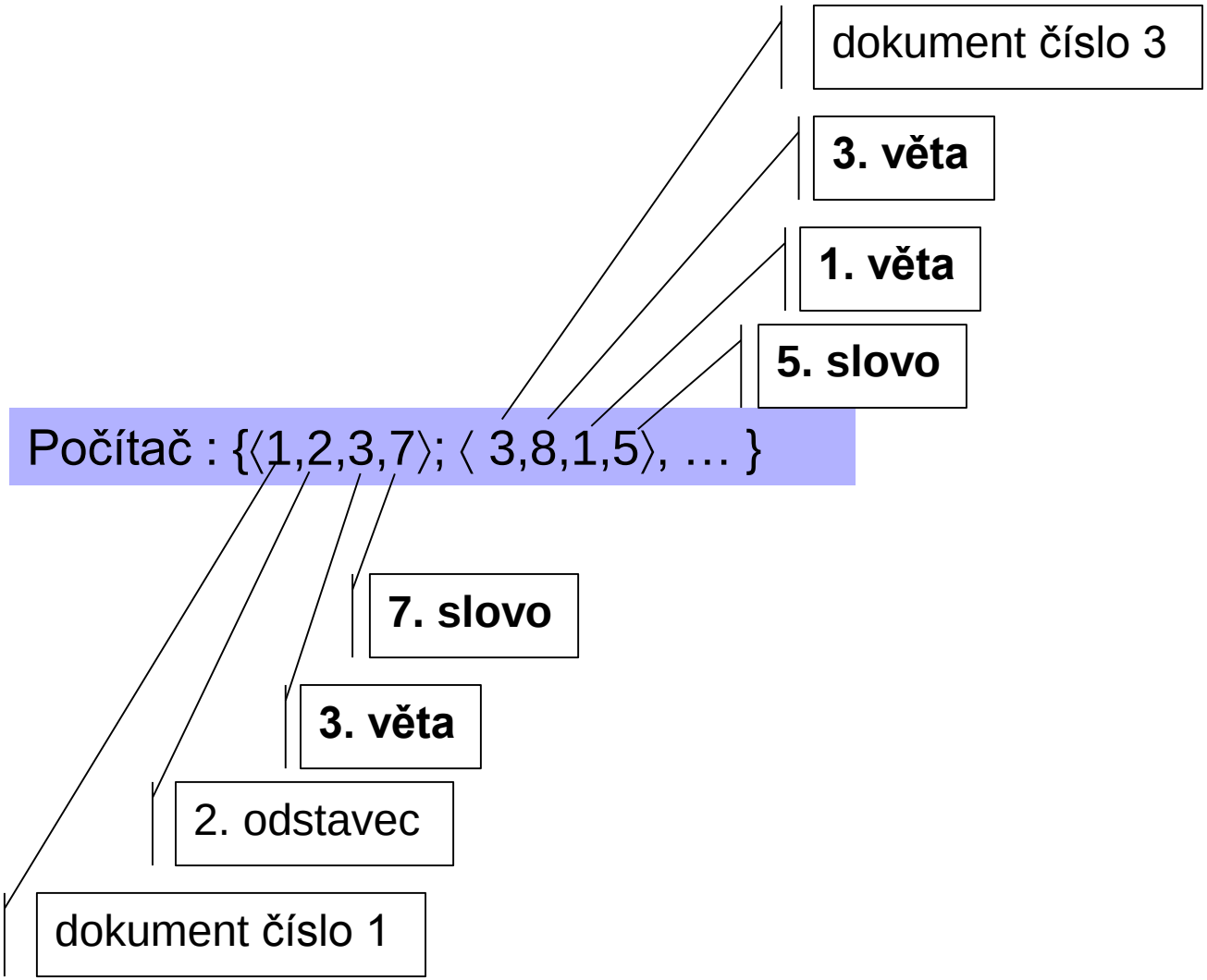
Dotaz: CD-ROM AND *ie

…	…
archeologie	1,3, 6, 8
…	…
biologie	2,4
…	…
chemie	3,7
…	…
filologie	3,5
…	…
filosofie	4, 6, 9
…	…
zoologie	1,2,8,9
…	…

Úsek invertovaného souboru se slovy „pozpátku“:

…	…
eifosolif	4, 6, 9
eigoloehcra	1,3, 6, 8
eigoloib	2,4
eigololif	3,5
eimehc	3,7
eiogolooz	1,2,8,9
…	…

Rozšíření invertovaného souboru o pozici slova



Následuje kvízová otázka

Kvíz – Otázka 2

Dokument
Klíčová slova

D1
Praha letiště fotbal

D2
Praha divadlo šachy

D3
Plzeň letiště fotbal

Která tabulka odpovídá invertovanému souboru pro dokumenty D1, D2 a D3?

Tabulka 1

divadlo	2,3
fotbal	1,3
letiště	1,3
Plzeň	3
Praha	1,2
šachy	2

Tabulka 2

divadlo	2
fotbal	3
letiště	1,3
Plzeň	3
Praha	1,2
šachy	2

Tabulka 3

divadlo	2
fotbal	1,3
letiště	1,3
Praha	1,2
Plzeň	3
šachy	2

Tabulka 4

divadlo	2
fotbal	1,3
letiště	1,3
Plzeň	3
Praha	1,2
šachy	2

Témata

- Vstupní zpracování
- Hodnocení úspěšnosti ukládání a vyhledávání
- Automatické indexování (automatická charakteristika obsahu)
- Booleovský model a jeho rozšiřování
- Invertovaný soubor
- Vektorový model

Vektorový model

- Východiska
- Principy
- Příklad – 5 dokumentů
- Frekvence termů a normalizované frekvence termů
- Inverzní dokumentová frekvence
- Schéma TF-IDF (Term Frequency – Inverse Document Frequency)
- Dotazy ve vektorovém modelu

Vektorový model – východiska

- Problémy s booleovským modelem
  - tvrdost boolevských operací, zejména AND
  - příliš mnoho nebo příliš málo výsledků
- Booleovský model vhodný pro experty, ne pro běžné uživatele
  - Většina neumí psát booleovské dotazy
  - I když umí, tak to považuje za ztrátu času
- Snaha po jednoduchosti
  - Dotaz – text v přirozeném jazyku nebo seznam slov
  - Výsledky setříděné podle relevance

Vektorový model – principy (1)

- Dokumenty D_1 , …, D_N
- Slova (termy) t_1,…,t_K
- Dokumenty reprezentovány pomocí vektorů vah termů
- $w_{i,j}$ je váha termu t_i v dokumentu D_j
- Dokument D_j lze chápat jako vektor $\langle w_{1,j}, \ldots, w_{K,j} \rangle$ vah termů

term	váha termu v dokumentu				
	D_1	…	D_j	…	D_N
t_1	$w_{1,1}$	…	$w_{1,j}$	…	$w_{1,N}$
…	…	…	…	…	…
t_i	$w_{i,1}$	…	$w_{i,j}$	…	$w_{i,N}$
…	…	…	…	…	…
t_K	$w_{K,1}$	…	$w_{K,j}$	…	$w_{K,N}$

Vektorový model – principy (2)

- Dotaz Q je reprezentován jako vektor $\langle q_1, ..., q_K \rangle$ vah termů
- Váha termu v dotazu vyjadřuje stupeň zajímavosti termu
- Počítá se míra podobnosti mezi dotazem a dokumentem jako míra podobnosti dvou vektorů
- Míra podobnosti se použije pro:
  - sestupné uspořádání vyhledaných dokumentů
  - omezení počtu dokumentů poskytovaných uživateli
  - optimalizaci dotazu na základě již nalezených podobných dokumentů

Příklad – 5 dokumentů

Za účelem srozumitelnosti vybrána pro indexaci pouze červená slova

- **D1 =** „Jestliže to chodí jako kachna a kváká jako kachna, tak to musí být kachna."
- **D2 =** „Pekingská kachna je oceňována zejména pro tenkou křupavou kachní kůži, která tvoří podstatnou část jídla. "
- **D3 =** „Bugyho vzestup mezi hvězdy přiměl animátory Warner studia přetvořit kachnu Daffy na silně závidějícího rivala králíka rozhodnutého získat zpět pozornost. Bugy si zatím nevšímá závisti kachny a nebo ji využívá ke své výhodě. To se ukázalo jako recept na úspěch tohoto dua."
- **D4 =** „18:25 26/3/2014 zápis v blogu: Našel jsem tento vynikající recept na králíka dušeného na víně na cookingforengineers.com.„
- **D5 =** „Li minulý týden ukázal jak dělat sečuánskou kachnu. Dnes budeme dělat čínské knedlíky (Jiaozi). Minulé léto jsem, měl šanci ochutnat toto jídlo v Pekingu. Je mnoho receptů na Jiaozi."

Frekvence termu v dokumentu

Váha $w_{i,j}$ termu t_i v dokumentu D_j = frekvence $f_{i,j}$ termu t_i v dokumentu D_j

term	frekvence termu v dokumentu				
	D_1	…	D_j	…	D_N
t_1	$f_{1,1}$	…	$f_{1,j}$	…	$f_{1,N}$
…	…	…	…	…	…
t_i	$f_{i,1}$	…	$f_{i,j}$	…	$f_{i,N}$
…	…	…	…	…	…
t_K	$f_{K,1}$	…	$f_{K,j}$	…	$f_{K,N}$

Frekvence termu v dokumentu – příklad

Příklad frekvencí pro termy z dokumentů D1, …, D5 (použita lematizace)

- **D1 =** „Jestliže to chodí jako kachna a kváká jako kachna, tak to musí být kachna."
- **D2 =** „Pekingská kachna je oceňována zejména pro tenkou křupavou kachní kůži, která tvoří podstatnou část jídla. "
- **D3 =** „Bugyho vzestup mezi hvězdy přiměl animátory Warner studia přetvořit kachnu Daffy na silně závidějícího rivala králíka rozhodnutého získat zpět pozornost. Bugy si zatím nevšímá závisti kachny a nebo ji využívá ke své výhodě. To se ukázalo jako recept na úspěch tohoto dua."
- **D4 =** „18:25 26/3/2014 zápis v blogu: Našel jsem tento vynikající recept na králíka dušeného na víně na cookingforengineers.com.„
- **D5 =** „Li minulý týden ukázal jak dělat sečuánskou kachnu. Dnes budeme dělat čínské knedlíky (Jiaozi). Minulé léto jsem, měl šanci ochutnat toto jídlo v Pekingu. Je mnoho receptů na Jiaozi."

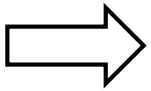
Term	frekvence termu v dokumentu				
	D1	D2	D3	D4	D5
jídlo		1			1
kachna	3	2	2		1
králík			1	1	
Peking		1			1
recept			1	1	1

Normalizovaná frekvence termu

Normalizovaná frekvence $tf_{i,j}$ termu t_i v dokumentu D_j:

$$tf_{i,j} = f_{i,j} / max_j \quad \text{kde} \quad max_j = max\{f_{1,j}, f_{2,j}, \ldots, f_{K,j}\}$$

term	frekvence termu v dokumentu				
	D_1	…	D_j	…	D_N
t_1	$f_{1,1}$	…	$f_{1,j}$	…	$f_{1,N}$
…	…	…	…	…	…
t_i	$f_{i,1}$	…	$f_{i,j}$	…	$f_{i,N}$
…	…	…	…	…	…
t_K	$f_{K,1}$	…	$f_{K,j}$	…	$f_{K,N}$
max	max_1	…	max_j	…	max_N



term	normalizovaná frekvence termu v dokumentu				
	D_1	…	D_j	…	D_N
t_1	$tf_{1,1}$	…	$tf_{1,j}$	…	$tf_{1,N}$
…	…	…	…	…	…
t_i	$tf_{i,1}$	…	$tf_{i,j}$	…	$tf_{i,N}$
…	…	…	…	…	…
t_K	$tf_{K,1}$	…	$tf_{K,j}$	…	$tf_{K,N}$

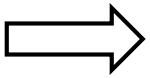
Cílem normalizace je aby váhy termů splňovaly $0 \leq tf_{i,j} \leq 1$

Normalizovaná frekvence termu – příklad

Normalizovaná frekvence $tf_{i,j}$ termu t_i v dokumentu D_j:

$$tf_{i,j} = f_{i,j} / max_j \quad \text{kde} \quad max_j = max\{f_{1,j}, f_{2,j}, \ldots, f_{K,j}\}$$

Term	frekvence termu v dokumentu				
	D1	D2	D3	D4	D5
jídlo		1			1
kachna	3	2	2		1
králík			1	1	
Peking		1			1
recept			1	1	1
max	3	2	2	1	1



Term	normalizované frekvence				
	D1	D2	D3	D4	D5
jídlo	0	0.5	0	0	1
kachna	1	1	1	0	1
králík	0	0	0.5	1	0
Peking	0	0.5	0	0	1
recept	0	0	0.5	1	1

Inverzní dokumentová frekvence

Inverzní dokumentová frekvence idf_i termu t_i v $\{D_1, \ldots, D_N\}$:

1. $df_{i,}$ = počet dokumentů, ve kterých se term t_i vyskytuje alespoň jednou
2. $idf_i = \log(\frac{N}{df_i})$

term	frekvence termu v dokumentu						
	D_1	…	D_j	…	D_N	df	idf
t_1	$f_{1,1}$	…	$f_{1,j}$	…	$f_{1,N}$	df_1	idf_1
…	…	…	…	…	…	…	…
t_i	$f_{i,1}$	…	$f_{i,j}$	…	$f_{i,N}$	df_i	idf_i
…	…	…	…	…	…	…	…
t_K	$f_{K,1}$	…	$f_{K,j}$	…	$f_{K,N}$	df_K	idf_K

Inverzní dokumentová frekvence termů – příklad

Inverzní dokumentová frekvence idf_i termu t_i v $\{D_1, \ldots, D_N\}$:

1. $df_{i,}$ = počet dokumentů, ve kterých se term t_i vyskytuje alespoň jednou
2. $idf_i = \log(\frac{N}{df_i})$

Term	frekvence termu v dokumentu						
	D1	D2	D3	D4	D5	*df*	*idf*
jídlo		1			1	2	0.398
kachna	3	2	2		1	4	0.097
králík			1	1		2	0.398
Peking		1			1	2	0.398
recept			1	1	1	3	0.222

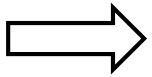
→ $\log(\frac{5}{df})$

TF-IDF váha termů

TF-IDF = Term Frequency - Inverse Document Frequency

TF-IDF váha $w_{i,j}$ termu t_i pro dokument D_j : $w_{i,j} = tf_{i,j} * idf_i$

term	normalizovaná frekvence termu					
	D_1	…	D_j	…	D_N	idf
t_1	$tf_{1,1}$	…	$tf_{1,j}$	…	$tf_{1,N}$	idf_1
…	…	…	…	…	…	…
t_i	$tf_{i,1}$	…	$tf_{i,j}$	…	$tf_{i,N}$	idf_i
…	…	…	…	…	…	…
t_K	$tf_{K,1}$	…	$tf_{K,j}$	…	$tf_{K,N}$	idf_K

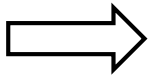


term	TF-IDF váha $w_{i,j}$ termu				
	D_1	…	D_j	…	D_N
t_1	$w_{1,1}$	…	$w_{1,j}$	…	$w_{1,N}$
…	…	…	…	…	…
t_i	$w_{i,1}$	…	$w_{i,j}$	…	$w_{i,N}$
…	…	…	…	…	…
t_K	$w_{K,1}$	…	$w_{K,j}$	…	$w_{K,N}$

TF-IDF váha termů - příklad

TF-IDF váha $w_{i,j}$ termu t_i pro dokument D_j : $w_{i,j} = tf_{i,j} * idf_i$

Term	normalizované frekvence					
	D1	D2	D3	D4	D5	*idf*
jídlo	0	0.5	0	0	1	0.398
kachna	1	1	1	0	1	0.097
králík	0	0	0.5	1	0	0.398
Peking	0	0.5	0	0	1	0.398
recept	0	0	0.5	1	1	0.222



term	TF-IDF váha $w_{i,j}$ termu				
	D1	D2	D3	D4	D5
jídlo	0.000	0.199	0.000	0.000	0.398
kachna	0.097	0.097	0.097	0.000	0.097
králík	0.000	0.199	0.000	0.398	0.000
Peking	0.000	0.199	0.000	0.000	0.398
recept	0.000	0.000	0.111	0.222	0.222

Dotazy ve vektorovém modelu

term	TF-IDF váha $w_{i,j}$ termu					dotaz
	D_1	…	D_j	…	D_N	Q
t_1	$w_{1,1}$	…	$w_{1,j}$	…	$w_{1,N}$	q_1
…	…	…	…	…	…	…
t_i	$w_{i,1}$	…	$w_{i,j}$	…	$w_{i,N}$	q_i
…	…	…	…	…	…	…
t_K	$w_{K,1}$	…	$w_{K,j}$	…	$w_{K,N}$	q_K

Dokument $D_j = \langle w_{1,j}, \ldots, w_{K,j} \rangle$

Dotaz $Q = \langle q_1, \ldots, q_K \rangle$

Vznik dotazu $Q = \langle q_1, \ldots, q_K \rangle$:

1. Uživatel může vyjádřit pomocí vah $\langle u_1, \ldots, u_K \rangle$ jak moc ho zajímají jednotlivé termy, nejjednoušší je $u_i = 1$ – zajímá a $u_i = 0$ – nezajímá
2. $q_i = u_i * idf_{i,j}$ pro i = 1, … K

Míra podobnosti dotazu a dokumentu

Nejčastěji používaná míra: kosinová míra podobnosti

Dokument D_j = $\langle w_{1,j},\ldots,w_{1,K}\rangle$ Dotaz Q = $\langle q_1,\ldots,q_K\rangle$

Kosinová míra podobnosti: $\frac{\sum_{i=1}^{K} w_{i,j} \times q_i}{\sqrt{\sum_{i=1}^{K} w_{i,j}^2} \times \sqrt{\sum_{i=1}^{K} q_i^2}}$

Míra podobnosti se použije pro:

- sestupné uspořádání vyhledaných dokumentů
- omezení počtu dokumentů poskytovaných uživateli
- optimalizaci dotazu na základě již nalezených podobných dokumentů

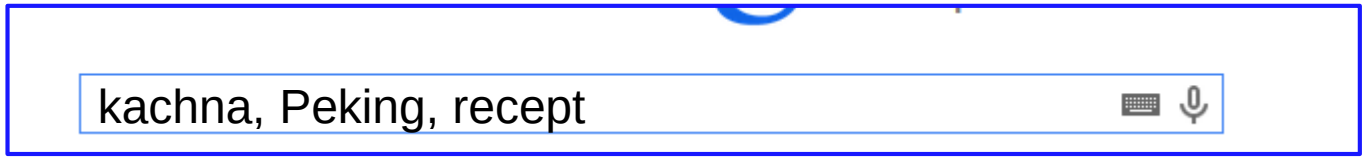
Míra podobnosti dotazu a dokumentu – příklady (1)

term	TF-IDF váha $w_{i,j}$ termu				
	D1	D2	D3	D4	D5
jídlo	0.000	0.199	0.000	0.000	0.398
kachna	0.097	0.097	0.097	0.000	0.097
králík	0.000	0.199	0.000	0.398	0.000
pečeně	0.000	0.000	0.000	0.000	0.000
Peking	0.000	0.199	0.000	0.000	0.398
recept	0.000	0.000	0.111	0.222	0.222

- **D1 =** „Jestliže to chodí jako kachna a kváká jako kachna, tak to musí být kachna."
- **D2 =** „Pekingská kachna je oceňována zejména pro tenkou křupavou kachní kůži, která tvoří podstatnou část jídla. "
- **D3 =** „Bugyho vzestup mezi hvězdy přiměl animátory Warner studia přetvořit kachnu Daffy na silně závidějícího rivala králíka rozhodnutého získat zpět pozornost. Bugy si zatím nevšímá závisti kachny a nebo ji využívá ke své výhodě. To se ukázalo jako recept na úspěch tohoto dua."
- **D4 =** „18:25 26/3/2014 zápis v blogu: Našel jsem tento vynikající recept na králíka dušeného na víně na cookingforengineers.com.„
- **D5 =** „Li minulý týden ukázal jak dělat sečuánskou kachnu. Dnes budeme dělat čínské knedlíky (Jiaozi). Minulé léto jsem, měl šanci ochutnat toto jídlo v Pekingu. Je mnoho receptů na Jiaozi."

termín, který nás také může zajímat

Zajímá nás recept na pekingskou kachnu:



Míra podobnosti dotazu a dokumentu – příklady (2)

Převod dotazu na vektor:

Term	u	idf	Q
jídlo	0	0.398	0.000
kachna	1	0.097	0.097
králík	0	0.398	0.000
pečeně	0	0.000	0.000
Peking	1	0.398	0.398
recept	1	0.222	0.222

- u – váha termu zadaná uživatelem
- idf – inverzní dokumentová frekvence
- $q_i = u * idf$ – váha termu v dotazu

$Q = \langle 0.000, 0.097, 0.000, 0.000, 0.398, 0.222 \rangle$

Míra podobnosti dotazu a dokumentu – příklady (3)

Kosinová míra podobnosti

D1
0.000
0.097
0.000
0.000
0.000
0.000

a

Q
0.000
0.097
0.000
0.000
0.398
0.222

:

$$\frac{\sum_{i=1}^{K} w_{i,j} \times q_i}{\sqrt{\sum_{i=1}^{K} w_{i,j}^2} \times \sqrt{\sum_{i=1}^{K} q_i^2}} = 0.208$$

Dokument	D1	D2	D3	D4	D5
Podobnost s Q	0.208	0.639	0.295	0.232	0.760
Pořadí	❺	❷	❸	❹	❶

Míra podobnosti dotazu a dokumentu – příklady (4)

Q: kachna, Peking, recept

Dokument	D1	D2	D3	D4	D5
Podobnost s Q	0.208	0.639	0.295	0.232	0.760
Pořadí	❺	❷	❸	❹	❶

❺
- **D1 =** „Jestliže to chodí jako kachna a kváká jako kachna, tak to musí být kachna."

❷
- **D2 =** „Pekingská kachna je oceňována zejména pro tenkou křupavou kachní kůži, která tvoří podstatnou část jídla. "

❸
- **D3 =** „Bugyho vzestup mezi hvězdy přiměl animátory Warner studia přetvořit kachnu Daffy na silně závidějícího rivala králíka rozhodnutého získat zpět pozornost. Bugy si zatím nevšímá závisti kachny a nebo ji využívá ke své výhodě. To se ukázalo jako recept na úspěch tohoto dua."

❹
- **D4 =** „18:25 26/3/2014 zápis v blogu: Našel jsem tento vynikající recept na králíka dušeného na víně na cookingforengineers.com.„

❶
- **D5 =** „Li minulý týden ukázal jak dělat sečuánskou kachnu. Dnes budeme dělat čínské knedlíky (Jiaozi). Minulé léto jsem, měl šanci ochutnat toto jídlo v Pekingu. Je mnoho receptů na Jiaozi."

Míra podobnosti dotazu a dokumentu – příklady (5)

Q: kachna, Peking, recept

5 ■ **D1 =** „Jestliže to chodí jako kachna a kváká jako kachna, tak to musí být kachna."

2 ■ **D2 =** „Pekingská kachna je oceňována zejména pro tenkou křupavou kachní kůži, která tvoří podstatnou část jídla. "

3 ■ **D3 =** „Bugyho vzestup mezi hvězdy přiměl animátory Warner studia přetvořit kachnu Daffy na silně závidějícího rivala králíka rozhodnutého získat zpět pozornost. Bugy si zatím nevšímá závisti kachny a nebo ji využívá ke své výhodě. To se ukázalo jako recept na úspěch tohoto dua."

4 ■ **D4 =** „18:25 26/3/2014 zápis v blogu: Našel jsem tento vynikající recept na králíka dušeného na víně na cookingforengineers.com.„

1 ■ **D5 =** „Li minulý týden ukázal jak dělat sečuánskou kachnu. Dnes budeme dělat čínské knedlíky (Jiaozi). Minulé léto jsem, měl šanci ochutnat toto jídlo v Pekingu. Je mnoho receptů na Jiaozi."

Dokument D5 je nejrelevantnější dotazu na recept na pekingskou kachnu. Ve skutečnosti se D5 ale týká sečuánské kachny, obecně jídla v Pekingu a receptů na knedlíky Jiaozi.

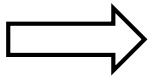
Míra podobnosti dotazu a dokumentu – příklady (6)

Pokus: Změna dokumentu D1 na
D1: „Jestliže to chodí jako kachna a kváká jako kachna z Pekingu, tak to musí být pekingská kachna."

místo

D1: „Jestliže to chodí jako kachna a kváká jako kachna, tak to musí být kachna."

Term	frekvence termu v dokumentu				
	D1	D2	D3	D4	D5
jídlo		1			1
kachna	3	2	2		1
králík			1	1	
Peking		1			1
recept			1	1	1



Term	frekvence termu v dokumentu				
	D1	D2	D3	D4	D5
jídlo		1			1
kachna	3	2	2		1
králík			1	1	
Peking	2	1			1
recept			1	1	1

Míra podobnosti dotazu a dokumentu – příklady (7)

Výsledek:

Q: kachna, Peking, recept ;

Dokument	D1	D2	D3	D4	D5
Podobnost s Q	0.727	0.480	0.485	0.281	0.605
Pořadí	❶	❹	❸	❺	❷

Původní

❶ ■ **D1 =** „Jestliže to chodí jako kachna a kváká jako kachna z Pekingu, tak to musí být pekingská kachna."

❹ ■ **D2 =** „Pekingská kachna je oceňována zejména pro tenkou křupavou kachní kůži, která tvoří podstatnou část jídla. "

❸ ■ **D3 =** „Bugyho vzestup mezi hvězdy přiměl animátory farner studia přetvořit kachnu Daffy na silně závidějícího rivala králíka rozhodnutého získat zpět pozornost. Bugy si zatím nevšímá závisti kachny a nebo ji využívá ke své výhodě. To se ukázalo jako recept na úspěch tohoto dua."

❺ ■ **D4 =** „18:25 26/3/2014 zápis v blogu: Našel jsem tento vynikající recept na králíka dušeného na víně na cookingforengineers.com.„

❷ ■ **D5 =** „Li minulý týden ukázal jak dělat sečuánskou kachnu. Dnes budeme dělat čínské knedlíky (Jiaozi). Minulé léto jsem, měl šanci ochutnat toto jídlo v Pekingu. Je mnoho receptů na Jiaozi."

❺ ❷ ❸ ❹ ❶