Metody pro zpracování nestrukturovaného textu: porovnávání ceníků

P. Hampl, H. Hamplová, J. Ivánek, R. Jiroušek, T. Kroupa, R. Lněnička, M. Studený, J. Vomlel

ÚTIA AV ČR a Empo, s.r.o

12. října 2005

Porovnávání ceníků

Předpokládáme následující scénář:

1. firma prodávající výpočetní techniku kupuje jednotlivé díly od různých dodavatelů
2. dodavatelé poskytují svoje ceníky
3. je výhodné porovnat cenu stejných komponent u různých dodavatelů (a např. si vybrat nejlevnějšího dodavatele)

Typický ceník

- obsahuje desetitisíce komponent
- je pouze částečně strukturovaný
- ceníky růných dodavatelů jsou vzájemně nekompatibilní

Porovnávání ceníků

Předpokládáme následující scénář:

1. firma prodávající výpočetní techniku kupuje jednotlivé díly od různých dodavatelů
2. dodavatelé poskytují svoje ceníky
3. je výhodné porovnat cenu stejných komponent u různých dodavatelů (a např. si vybrat nejlevnějšího dodavatele)

Typický ceník

- obsahuje desetitisíce komponent
- je pouze částečně strukturovaný
- ceníky různých dodavatelů jsou vzájemně nekompatibilní

Detailní specifikace ceníku

Ceník je tabulka, ve které

- řádky jsou počítačové komponenty
- sloupce jsou atributy komponent

Atributy

- jednoznačný identifikátor komponenty - part number (pouze někdy)
- výrobce komponenty (často pouze zkratka)
- kategorie a případně podkategorie, do které komponenta patří
- popis komponenty
- cena komponenty

Detailní specifikace ceníku

Ceník je tabulka, ve které

- řádky jsou počítačové komponenty
- sloupce jsou atributy komponent

Atributy

- jednoznačný identifikátor komponenty - part number (pouze někdy)
- výrobce komponenty (často pouze zkratka)
- kategorie a případně podkategorie, do které komponenta patří
- popis komponenty
- cena komponenty

Příklad dvou ceníků

L_1	
Popis	*Cena*
Cisco 828 G.SHDSL Router 1E, 1G.SHDSL	13 640
HP LaserJet 3030 Print/Scan/Copy/Fax, Paralel,USB	12 529
HP PL DL360R04 X3.0/1M 1G SA6i iLO	59 453
...	...

L_2	
Popis	*Cena*
Cisco 828 (G.SHDSL)	13 740
LaserJet 3030, tiskárna, kopírka, skener, fax	12 199
...	...

Úloha

Pro zadaný popis D komponenty nalézt v cenících $L_1, \ldots, L_k$ popisy komponent $D_1, \ldots, D_k$, které se nejvíce podobají popisu D.

Příklad výstupu

Popis D	*cena v L_1*	*cena v L_2*
Cisco 828 G.SHDSL router	13 640	13 740
HP LaserJet 3030 All-in-One	12 529	12 199
HP PL DL360R04 X3.0/1M 1G SA6i iLO	59 453	-
...	...	...

Úloha

Pro zadaný popis D komponenty nalézt v cenících $L_1, \ldots, L_k$ popisy komponent $D_1, \ldots, D_k$, které se nejvíce podobají popisu D.

Příklad výstupu

Popis D	*cena v* L_1	*cena v* L_2
`Cisco 828 G.SHDSL router`	13 640	13 740
`HP LaserJet 3030 All-in-One`	12 529	12 199
`HP PL DL360R04 X3.0/1M 1G SA6i iLO`	59 453	-
...	...	...

Porovnávání řetězců

- Komponenty jsou k sobě přiřazeny na základě podobnosti jejich popisu.
- Popis je řetězec znaků.

Dva základní přístupy pro porovnávání dvou řetězců

1. techniky založené na znakovém přístupu
   - podobnost je počítána na znakové úrovni
   - příkladem metody je String Edit Distance - kdy cílem je minimalizace celkové ceny při změnách jednoho popisu na druhý pomocí operací jako např. vkládání znaků, vymazání znaku, či změna znaku.
2. techniky založené na vektorovém přístupu
   - řetězec je rozdělen na jednotlivá slova - tokens
   - podobnost se počítá na základě výskytu tokens v obou porovnávaných řetězcích

Porovnávání řetězců

- Komponenty jsou k sobě přiřazeny na základě podobnosti jejich popisu.
- Popis je řetězec znaků.

Dva základní přístupy pro porovnávání dvou řetězců

1. techniky založené na znakovém přístupu
   - podobnost je počítána na znakové úrovni
   - příkladem metody je String Edit Distance - kdy cílem je minimalizace celkové ceny při změnách jednoho popisu na druhý pomocí operací jako např. vkládání znaků, vymazání znaku, či změna znaku.
2. techniky založené na vektorovém přístupu
   - řetězec je rozdělen na jednotlivá slova - tokens
   - podobnost se počítá na základě výskytu tokens v obou porovnávaných řetězcích

Porovnávání řetězců

- Komponenty jsou k sobě přiřazeny na základě podobnosti jejich popisu.
- Popis je řetězec znaků.

Dva základní přístupy pro porovnávání dvou řetězců

1. techniky založené na znakovém přístupu
   - podobnost je počítána na znakové úrovni
   - příkladem metody je String Edit Distance - kdy cílem je minimalizace celkové ceny při změnách jednoho popisu na druhý pomocí operací jako např. vkládání znaků, vymazání znaku, či změna znaku.
2. techniky založené na vektorovém přístupu
   - řetězec je rozdělen na jednotlivá slova - tokens
   - podobnost se počítá na základě výskytu tokens v obou porovnávaných řetězcích

Naše definice podobnosti

Součet délek podřetězců prvního řetězce obsažených v druhém řetězci a dlouhých minimálně dva znaky.

Definition

$$Similarity(R_1, R_2) = \sum_{k=1}^{Length(R_1)-1} \sum_{\substack{R \in \mathcal{R}_1^k \cap \mathcal{R}_2 : \\ Length(R) \geq 2}} Length(R)$$

- $\mathcal{R}_i^k$ je množina všech podřetězců řetězce R_i začínajících na k-té pozici
- $\mathcal{R}_i$ je množina všech podřetězců řetězce R_i

Příklad výpočtu podobnosti

R_1

W	I	N	D	O	W	S		T	E	R	M

R_2

W	I	N		T	R	M	N	L

Příklad výpočtu podobnosti

R_1

W	I	N	D	O	W	S		T	E	R	M

R_2

W	I	N		T	R	M	N	L

$k = 1$

$Similarity(R_1, R_2) = 0$

Příklad výpočtu podobnosti

R_1

W	I	N	D	O	W	S		T	E	R	M

R_2

W	I	N		T	R	M	N	L

$k = 1$

$R =$"WI"

$Length(R) = 2$

$Similarity(R_1, R_2) = 2$

Příklad výpočtu podobnosti

R_1

W	I	N	D	O	W	S		T	E	R	M

R_2

W	I	N		T	R	M	N	L

$k = 1$

$R =$"WIN"

$Length(R) = 3$

$$Similarity(R_1, R_2) = 2 + 3$$

Příklad výpočtu podobnosti

R_1

W	I	N	D	O	W	S		T	E	R	M

R_2

W	I	N		T	R	M	N	L

$k = 2$

$Similarity(R_1, R_2) = 2 + 3$

Příklad výpočtu podobnosti

R_1

W	I	N	D	O	W	S		T	E	R	M

R_2

W	I	N		T	R	M	N	L

$k = 2$

$R =$"IN"

$Length(R) = 2$

$$Similarity(R_1, R_2) = 2 + 3 + 2$$

Příklad výpočtu podobnosti

R_1

W	I	**N**	D	O	W	S		T	E	R	M

R_2

W	I	N		T	R	M	N	L

$k = 3$

$$Similarity(R_1, R_2) = 2 + 3 + 2$$

Příklad výpočtu podobnosti

R_1

W	I	N	**D**	O	W	S		T	E	R	M

R_2

W	I	N		T	R	M	N	L

$k = 4$

$$Similarity(R_1, R_2) = 2 + 3 + 2$$

Příklad výpočtu podobnosti

R_1

W	I	N	D	**O**	W	S		T	E	R	M

R_2

W	I	N		T	R	M	N	L

$k = 5$

$$Similarity(R_1, R_2) = 2 + 3 + 2$$

Příklad výpočtu podobnosti

R_1

W	I	N	D	O	**W**	S		T	E	R	M

R_2

W	I	N		T	R	M	N	L

$k = 6$

$$Similarity(R_1, R_2) = 2 + 3 + 2$$

Příklad výpočtu podobnosti

R_1

W	I	N	D	O	W	**S**		T	E	R	M

R_2

W	I	N		T	R	M	N	L

$k = 7$

$$Similarity(R_1, R_2) = 2 + 3 + 2$$

Příklad výpočtu podobnosti

R_1

W	I	N	D	O	W	S	_	T	E	R	M

R_2

W	I	N		T	R	M	N	L

$k = 8$

$$Similarity(R_1, R_2) = 2 + 3 + 2$$

Příklad výpočtu podobnosti

R_1

W	I	N	D	O	W	S	_	T	E	R	M

R_2

W	I	N	_	T	R	M	N	L

$k = 8$

$R =$” T”

$Length(R) = 2$

$$Similarity(R_1, R_2) = 2 + 3 + 2 + 2$$

Příklad výpočtu podobnosti

R_1

W	I	N	D	O	W	S		**T**	E	R	M

R_2

W	I	N		T	R	M	N	L

$k = 9$

$$Similarity(R_1, R_2) = 2 + 3 + 2 + 2$$

Příklad výpočtu podobnosti

R_1

W	I	N	D	O	W	S		T	**E**	R	M

R_2

W	I	N		T	R	M	N	L

$k = 10$

$$Similarity(R_1, R_2) = 2 + 3 + 2 + 2$$

Příklad výpočtu podobnosti

R_1

W	I	N	D	O	W	S		T	E	**R**	M

R_2

W	I	N		T	R	M	N	L

$k = 11$

$$Similarity(R_1, R_2) = 2 + 3 + 2 + 2$$

Příklad výpočtu podobnosti

R_1

W	I	N	D	O	W	S		T	E	**R**	M

R_2

W	I	N		T	R	M	N	L

$k = 11$

$R =$”RM”

$Length(R) = 2$

$$Similarity(R_1, R_2) = 2 + 3 + 2 + 2 + 2$$

Příklad výpočtu podobnosti

R_1

W	I	N	D	O	W	S		T	E	R	M

R_2

W	I	N		T	R	M	N	L

$$Similarity(R_1, R_2) = 2 + 3 + 2 + 2 + 2 = 11$$

Výsledky

- 5 ceníků - jeden z nich byl referenční
- testovací data: 277 komponent z referenčního ceníku

Hodnocení

- přesnost = počet správně nalezených komponent / celkový počet komponent v testovacích datech
- TLE = počet správně nalezených komponent / počet komponent s vyšší než prahovou hodnotou podobnosti
- PNM = počet správně nalezených komponent se *stejným part number* / celkový počet komponent v testovacích datech

Výsledky (pokrač.)

ceník	*počet položek*	*přesnost*	*TLE*	*PNM*
2	9 670	19,49%	81,25%	9,75%
3	7 941	19,49%	86,84%	13,72%
4	24 076	55,96%	92,05%	33,94%
5	22 182	40,43%	94,74%	23,83%

Problémy přístupu:

- při pouhé permutaci slov v popisu se výsledná podobnost výrazně sníží
- není možné uvažovat zkratky, synonyma, termíny v jiném jazyce protože nepracujeme se slovy, ale se znaky

Vektorový přístup

Definition

$$Similarity(R_1, R_2) = \sum_{i=1}^{d} w(x_i, R_1) \cdot w(x_i, R_2)$$

- d je počet slov (tokens)
- $w(x_i, R_j)$ je normalizovaná váha slova (token) v řetězci R_j. Pokud se slovo v řetězci nevyskytuje je rovna nule.

Výsledky

ceník	*přesnost*	*přesnost po vyřazení neexist. položek*
2	20,95%	37,35%
3	16,89%	34,27%
4	40,54%	46,51%
5	36,49%	52,94%

Problémy:

- jak rozdělit řetězec na slova (jaké znaky použít jako separátory)?
- je třeba vytvořit slovník synonym, zkratek, případně i významů slov.

Co bychom chtěli v budoucnu udělat:

- vytvořit slovník synonym, zkratek - nejlépe i s významy slov
  - např. výrobce, název komponenty, parametr komponenty
- využít částečnou strukturu popisu komponenty rozdělenou do atributů
  - v některých cenících jsou atributy: výrobce, skupina komponent
- použít jednoduchá produkční pravidla (bezkontextové či regulární gramatiky) pro definici některých významových skupin
  - např. 1,2 GHz