dnevi slovenske informatike

"Ustvarimo nove rešitve!"

Kongresni center Grand hotel Bernardin
Portorož | Slovenija | 16. - 18. april 2012

19. konferenca
Dnevi slovenske informatike

POVEZOVANJE KRATKIH ZAPISOV S TIPOGRAFSKIMI NAPAKAMI – PRIMER APLIKACIJE ALGORITMA EDIT DISTANCE

Uroš Godnov
UP, Fakulteta za management

17. 04. 2012

Tomaž Dular
Mit inženiring d.o.o.

slovensko društvo informatika

Agenda

- Problematika kakovosti področja podatkov
- Ožje problemsko področje
- Predstavitev algoritma
- Kreiranje in ustvarjanje algoritma
- Rezultati simulacije
- Zaključek.

Širše problemsko področje

Visoki stroški

- **All Quiet on the Western Front**
- Približno 600 milijard dolarjev stroškov letno v ZDA
- Veliko govorjenja o kakovosti podatkov, ni pa modela prihrankov in stroškov upravljanja s kakovostjo podatkov

Gartner 2010:

- povprečna anketirana organizacija letno 8.2 milijonov stroškov zaradi nekakovostnih podatkov
- 22 % anketiranih organizacij ima letno 20 milijonov dolarjev stroškov zaradi nekakovostnih podatkov
- 4 % anketiranih organizacij letno več kot 100 milijonov stroškov zaradi nekakovostnih podatkov

Ožje problemsko področje

Levji delež dela

- Standardizacija, dopolnitve ter popravki terjajo največ časa za izgradnjo analitičnih sistemov
- Raziskave so pokazale, da 90 % časa potrebujemo za odstranjevanje podvojenih zapisov (Fayyad in drugi)
- Problematična je tudi integracija podatkov (fokus našega članka)

Tipografske napake:

- Janez Novak → Novak Janez
- Janez Novak → Janez Novka
- Janez Novak → g. Janez Novak
- Janez Novak → g. Novka Janez

Ožje problemsko področje

Malce zgodovine → delovno intenzivni pristop

- Potreba po povezovanju je posledica decentralizacije podatkov.
- Leta 1990 Decennial Census → 3000 uradnikov, 3 mesece
- Kanadski statistični urad na področju kmetijstva: pred letom 1992 → 75 uradnikov, 3 mesece

Računalniški pristop:

- prvi primer: 6 tednov
- drugi primer 6500 ur

Ožje problemsko področje

Zakaj razlike v zapisih v iz različnih informacijskih virov

- Različna skalarnost zapisov
- Način shranjevanja podatkov
- Različne krajšave zapisov
- Tipografske napake

Povezovanje iz kakovostnih informacijskih virov (Winkler):

- 20+ % napak pri povezovanju imen
- 10+ % napak pri povezovanju priimkov

Računalniško povezovanje in deduplikacija podatkov

Različni algoritmi

- Prisotnost/odsotnost tipografskih napak
- Prisotnost/odsotnost vedenja, kako se tipografske napake pojavljajo

Prisotnost tipografskih napak, odsotnost poznavanja vzorca pojavljanja tipografskih napak

- **Jaro-Winkler algoritem**
- **Edit distance algoritem**
- Jaccard index
- Simil index
- Hibridne izvedenke

Edit distance algoritem

Klasični Edit distance algoritem

- Najmanjše število potrebnih operacij (vstavi ter briši), da en zapis spremenimo v drugega

		V	O	D	N	O	V
	0	1	2	3	4	5	6
G	1	1	2	3	4	5	6
O	2	2	1	2	4	4	6
D	3	3	2	1	2	3	4
N	4	4	3	2	1	2	3
O	5	5	4	3	2	1	2
V	6	6	5	4	3	2	**1**

Edit distance algoritem

Hibridni Edit distance algoritem

$$sim(s, t)=\frac{1}{K}\sum_{i=1}^{K} max_{j=1}^{L}\ SIM\ (A_i, B_j)$$

GODNOV			VODNOV		
Dolžina podnizov					
3	4	5	3	4	5
GOD	GODN	GODNO	OD	VODN	VODNO
ODN	ODNO	ODNOV	ODN	ODNO	ODNOV
DNO	DNOV		DNO	DNOV	
NOV			NOV		

$$sim(GODNOV, VODNOV)=\frac{1}{4}(0.67 + 1 + 1 + 1)=0.915$$

Simulacija

Zbirka imen in priimkov

- Približno 18 000 zapisov iz demonstracijske zbirke podatkov AdventureWorks
- 6592 smo jih okvarili
- Okvare so bile slučajno porazdeljene, okvare v prvi polovici zapisa ter okvare v drugi polovici zapisa
- min dolžina – 6; max dolžina – 20; povp. dolžina - 12

Primer okvar

Correct record	Random errors	Errors in the first half	Errors in the second half
Cassie Chande	Cansie Chande	Cansie Chande	Cassie Cnande
Edgar Sara	gdgjr Sara	jdggr Sara	EdgarjSaga
Candace Fernandez	Candace Ferrancez	Canranc Fernandez	Candace Ferrancez

Simulacija

Simulacija klasičnega in hibridnega Edit distance algoritma

- Iskanje optimalne dolžine podnizov za hibridni algoritem
- Vzorec najboljšega oz. najvišjega indeksa vrednosti algoritma

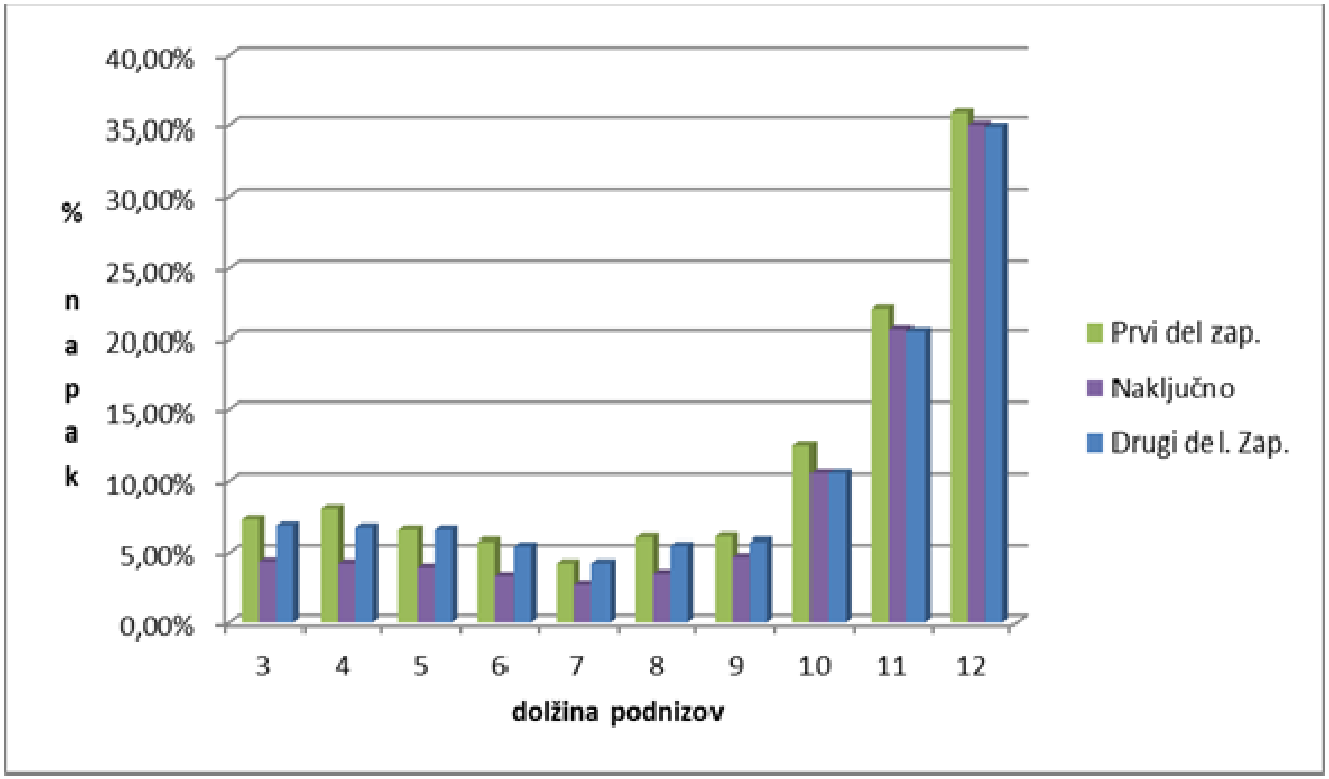
		Dolžina zapisa														
	Algoritem	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Edit distance	**Klasičen**	0 6667	0,7875	0,7942	0,8256	0,8379	0,8493	0,8598	0,8704	0,8772	0,8828	0,8864	0,9017	0,8955	0,8967	0,9071
	Hibridni - dol. podnizov 3	0,7500	0,8184	0,8141	0,8348	0,8403	0,8504	0,8581	0,8711	0,8779	0,8824	0,8871	0,9013	0,8949	0,8963	0,9048
	Hibridni - dol. podnizov 4	0,5625	0,8171	0,8097	0,8428	0,8450	0,8550	0,8617	0,8729	0,8788	0,8824	0,8878	0,9013	0,8936	0,8936	0,9048
	Hibridni - dol. podnizov 5	0,6500	0,8138	0,8000	0,8392	0,8510	0,8611	0,8656	0,8757	0,8809	0,8837	0,8884	0,9028	0,8940	0,8970	0,9082
	Hibridni - dol. podnizov 6	0,6667	0,8028	0,8006	0,8391	0,8512	0,8600	0,8697	0,8783	0,8827	0,8856	0,8889	0,9042	0,8955	0,8968	0,9084
	Hibridni - dol. podnizov 7		0,7875	0,7988	0,8341	0,8456	0,8595	0,8693	0,8778	0,8850	0,8873	0,8913	0,9054	0,8978	0,8968	0,9099

Simulacija

Mikroanaliza → najvišji indeks algoritma

- Računanje indeksa je pokazala, da je optimalna dolžina podniza min(s,t) ter min(s,t)+1

Makroanaliza → število napak pri povezovanju zapisov



Simulacija

Rezultati simulacije

		Značilnosti zapisov		
		Napake v prvi pol.	Napake v drugi pol.	Naključno porazdeljene napake
algoritem				
Edit distance	Klasičen	3,67%	2,87%	2,55%
	hibridni $D_1=dol_{MIN}$ (s, t)/2	3,20%	3,03%	1,97%
	hibridni $D_2=dol_{MIN}$ (s, t)/2 +1	2,42%	2,72%	1,66%

Zaključek

- Hibridni Edit distance algoritem z dolžino podniza min(s,t)+1 je učinkovitejši od regularnega Edit distance algoritma in prav tako ni občutljiv na pojavljanje mesta tipografskih napak.

Hvala za vašo pozornost !

Vprašanja?

Pripombe?

Predlogi?