

Intelligentno pridobivanje podatkov iz poljubnih virov

Slavko Žitnik

Mentor: izr. prof. dr. Marko Bajec

Fakulteta za računalništvo in informatiko
Fakulteta za matematiko in fiziko

September 2010

Agenda

- 1 Zgodovina
- 2 Sistem iCORE
- 3 Komponente sistema iCORE
- 4 Vprašanja
- 5 *Dodatno*

Zgodovina

1945: “As We May Think”, Vannevar Bush

1960: SMART, Gerard Salton

Search 1.0 (1996): ujemanje besed

Search 2.0 (1999): SEO, SEM

Search 3.0 (2007): horizontalno, vertikalno

Search 4.0 (2009): personalizacija, socialna omrežja

Zgodovina

1945: “As We May Think”, Vannevar Bush

1960: SMART, Gerard Salton

Search 1.0 (1996): ujemanje besed

Search 2.0 (1999): SEO, SEM

Search 3.0 (2007): horizontalno, vertikalno

Search 4.0 (2009): personalizacija, socialna omrežja

Zgodovina

1945: “As We May Think”, Vannevar Bush

1960: SMART, Gerard Salton

Search 1.0 (1996): **ujemanje besed**

Search 2.0 (1999): SEO, SEM

Search 3.0 (2007): horizontalno, vertikalno

Search 4.0 (2009): personalizacija, socialna omrežja

Zgodovina

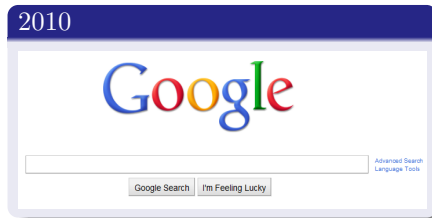
Search 1.0 (1996): ujemanje besed

Search 2.0 (1999): **SEO, SEM**

Search 3.0 (2007): horizontalno, vertikalno

Search 4.0 (2009): personalizacija, socialna omrežja

Search 5.0 (2011): iCORE - intelligent context-aware (multisource)
information retrieval



Zgodovina

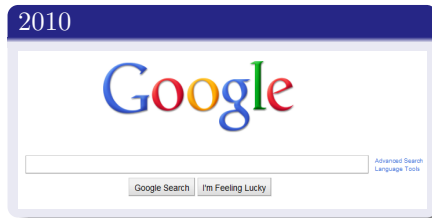
Search 1.0 (1996): ujemanje besed

Search 2.0 (1999): SEO, SEM

Search 3.0 (2007): **horizontalno, vertikalno**

Search 4.0 (2009): personalizacija, socialna omrežja

Search 5.0 (2011): iCORE - intelligent context-aware (multisource)
information retrieval



Zgodovina

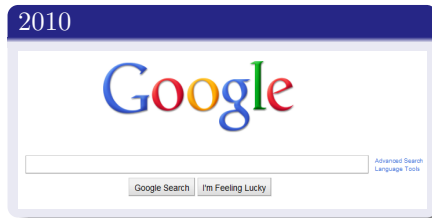
Search 1.0 (1996): ujemanje besed

Search 2.0 (1999): SEO, SEM

Search 3.0 (2007): horizontalno, vertikalno

Search 4.0 (2009): **personalizacija, socialna omrežja**

Search 5.0 (2011): iCORE - intelligent context-aware (multisource)
information retrieval



Zgodovina

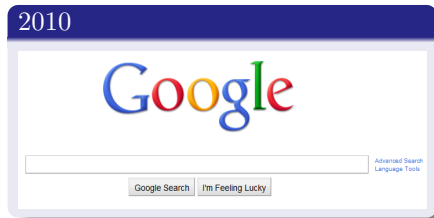
Search 1.0 (1996): ujemanje besed

Search 2.0 (1999): SEO, SEM

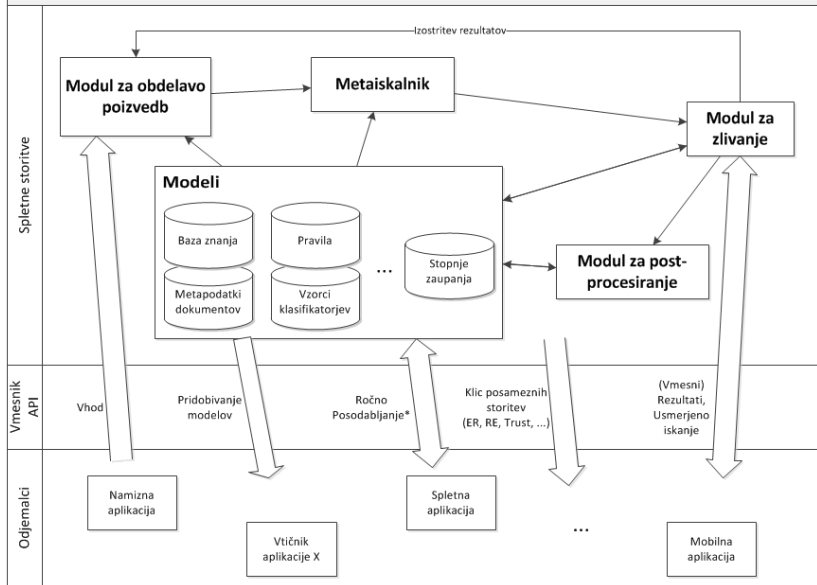
Search 3.0 (2007): horizontalno, vertikalno

Search 4.0 (2009): personalizacija, socialna omrežja

Search 5.0 (2011): **iCORE - intelligent context-aware (multisource)
information retrieval**

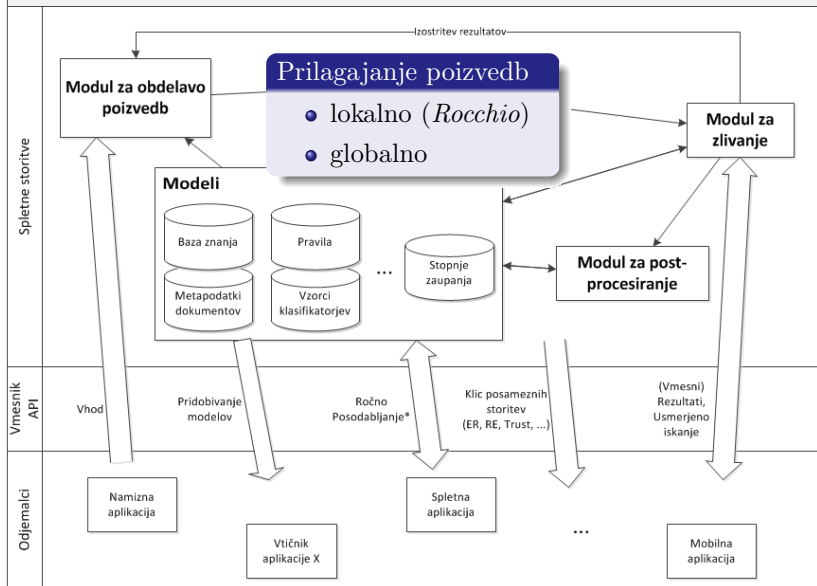


Sistem iCORE



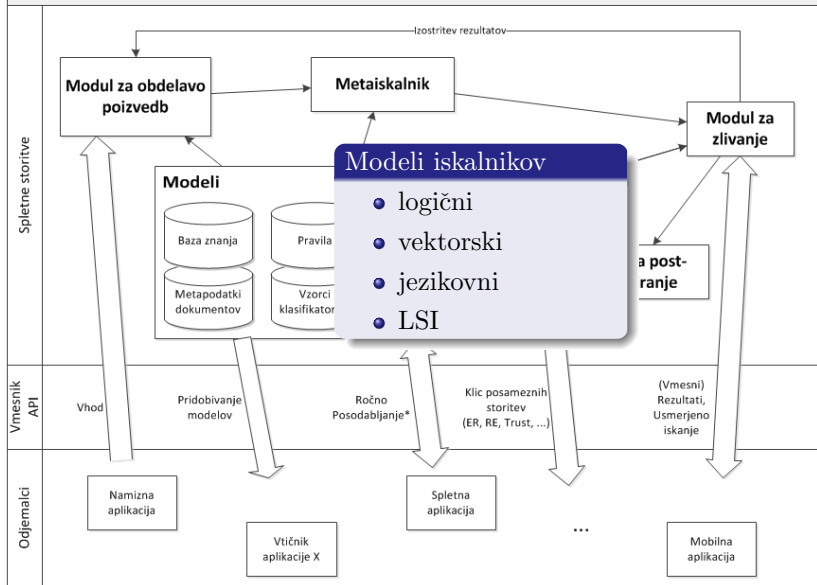
*Opcijsko, če avtomatsko ni dovolj natančno

Sistem iCORE



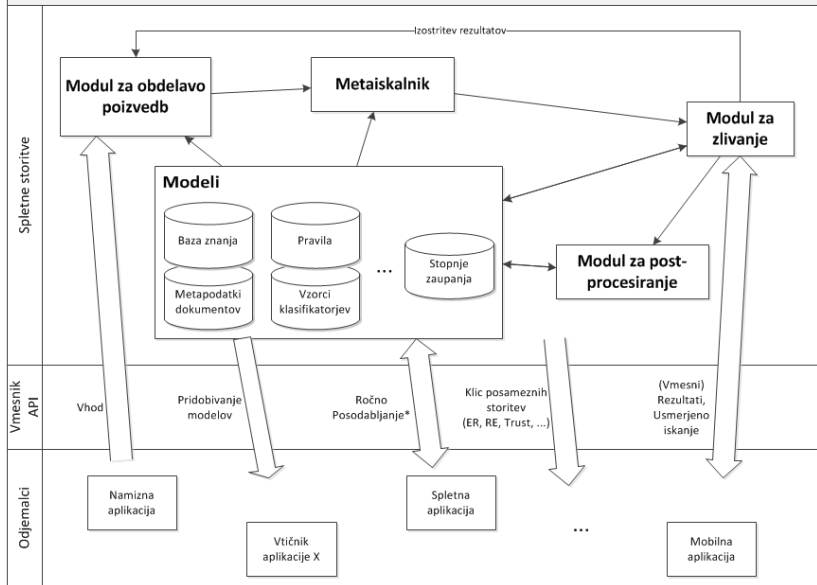
*Opcijsko, če avtomatsko ni dovolj natančno

Sistem iCORE



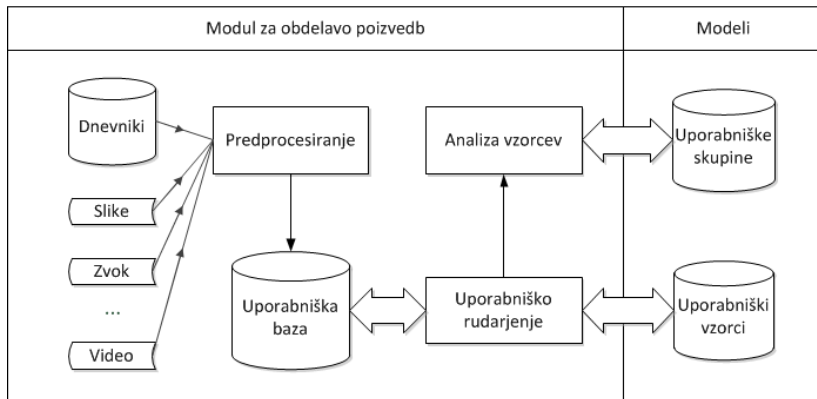
*Opcijsko, če avtomatsko ni dovolj natančno

Sistem iCORE

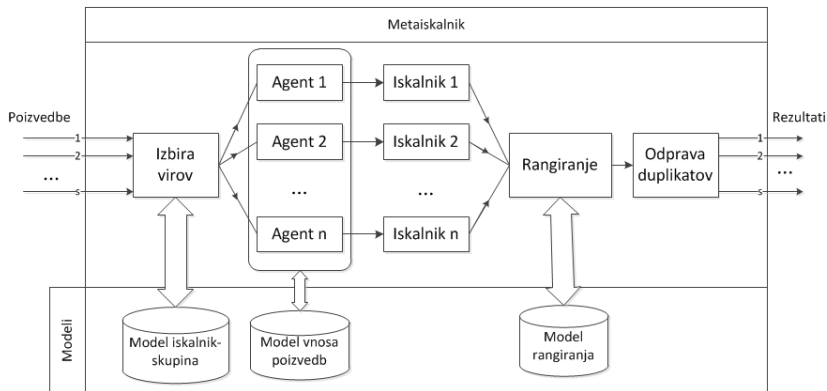


*Opcijsko, če avtomatsko ni dovolj natančno

Modul za obdelavo poizvedb



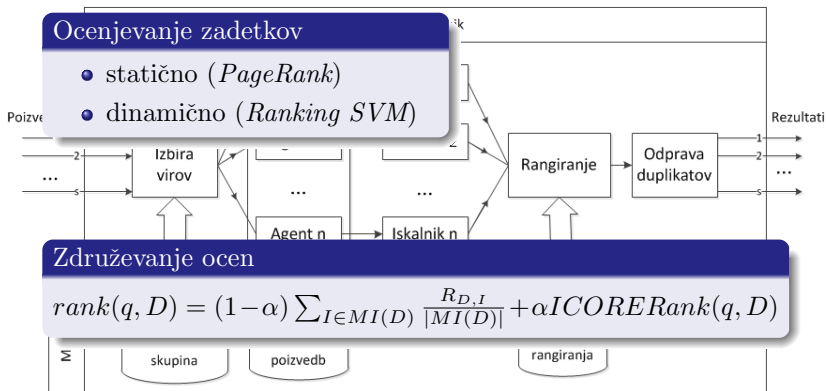
Metaiskalnik



Metaiskalnik

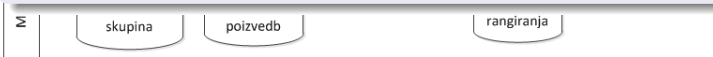
Ocenjevanje zadetkov

- statično (*PageRank*)
- dinamično (*Ranking SVM*)

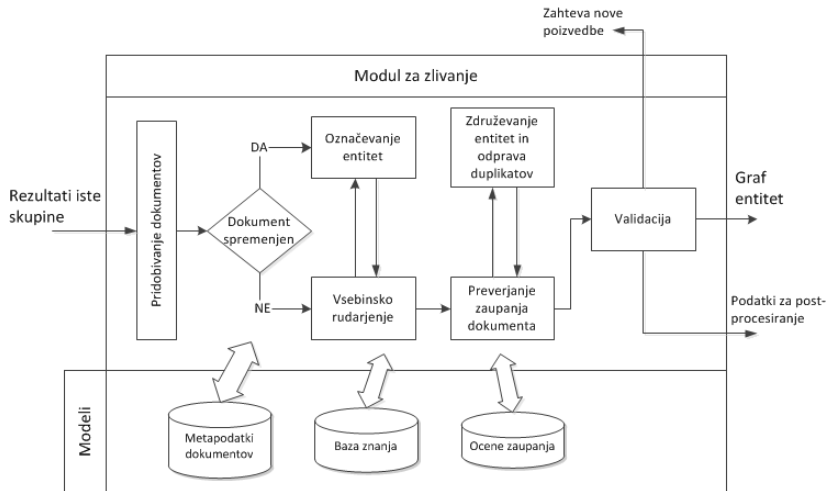


Združevanje ocen

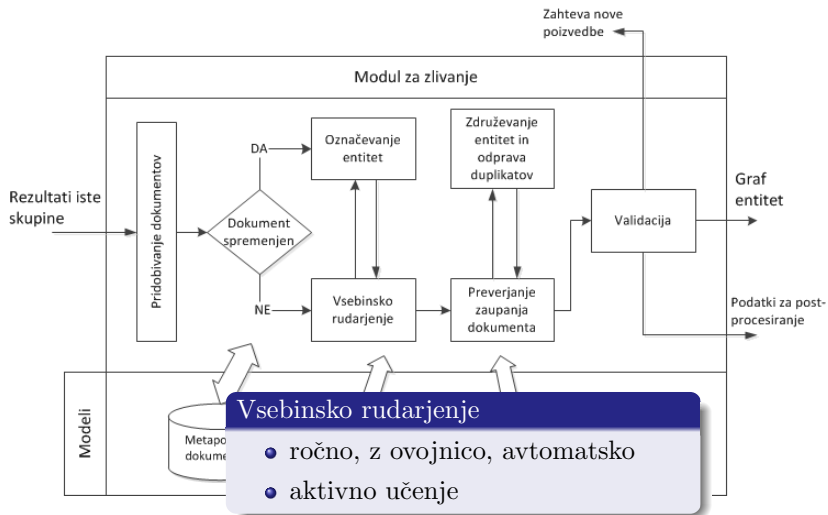
$$\text{rank}(q, D) = (1 - \alpha) \sum_{I \in MI(D)} \frac{R_{D,I}}{|MI(D)|} + \alpha \text{ICORERank}(q, D)$$



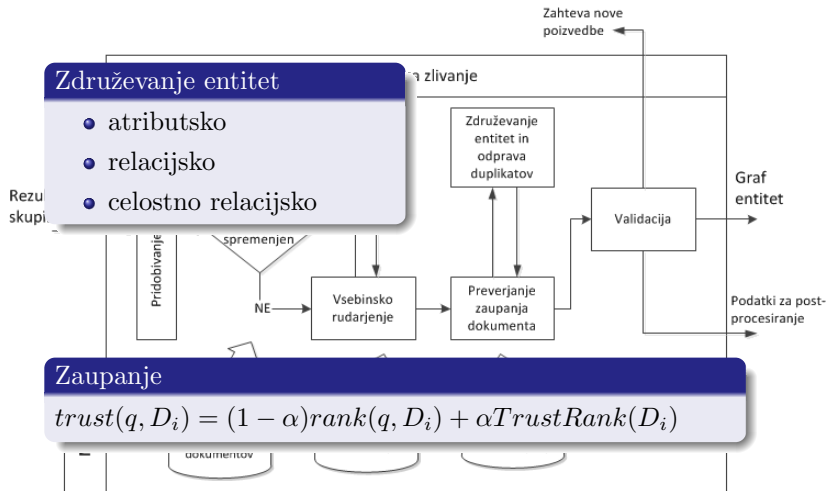
Modul za zlivanje



Modul za zlivanje



Modul za zlivanje

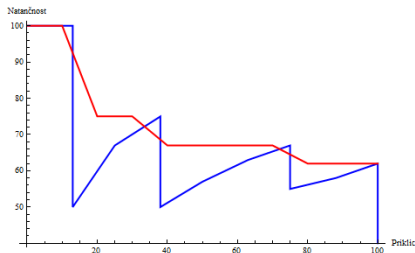


Modul za post-procesiranje in modeli

- zajem testnih in učnih podatkov
 - pridobitev uporabnikovega odziva
 - posodabljanje modelov (klasifikatorji, baze znanja)
 - odloženo procesiranje s skladom
 - ročno vnašanje sprememb
-
- dve bazi znanja
 - metapodatki dokumentov, ocene zaupanja
 - vzorci klasifikatorjev za ocenjevanje, obdelavo poizvedb, kontekst uporabnika

Evaluacija

- ocenjevanje sistema
- ocenjevanje komponent
 - testni nabori podatkov
 - natančnost
 - priklic
 - F ocena
 - ocenjevanje rangiranja



Slika: Krivulja Natančnost-priklic za podatke iz diplomskega dela

Vprašanja?

Gerard Salton Award



Gerard Salton
1983



Karen Spaerck Jones
1988



Cyril Cleverdon
1991



William Cooper
1994



Tefko Saracevic
1997



Stephen Robertson
2000



W. Bruce Croft
2003

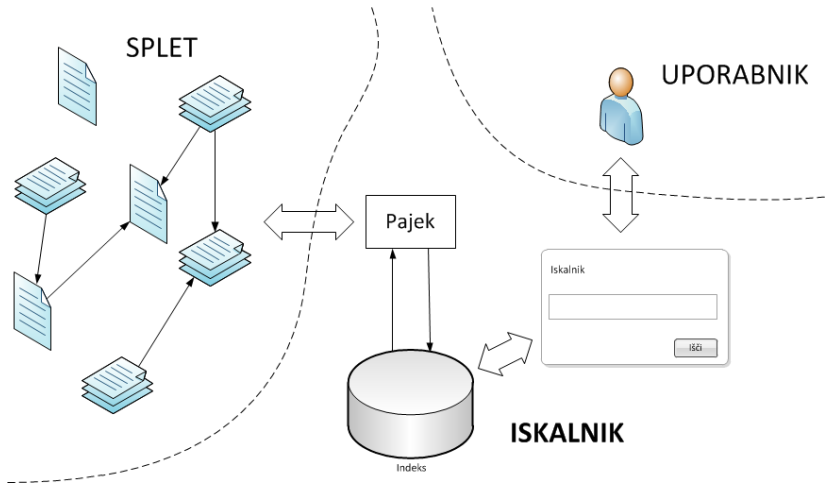


C.J. "Keith" van Rijsbergen
2006

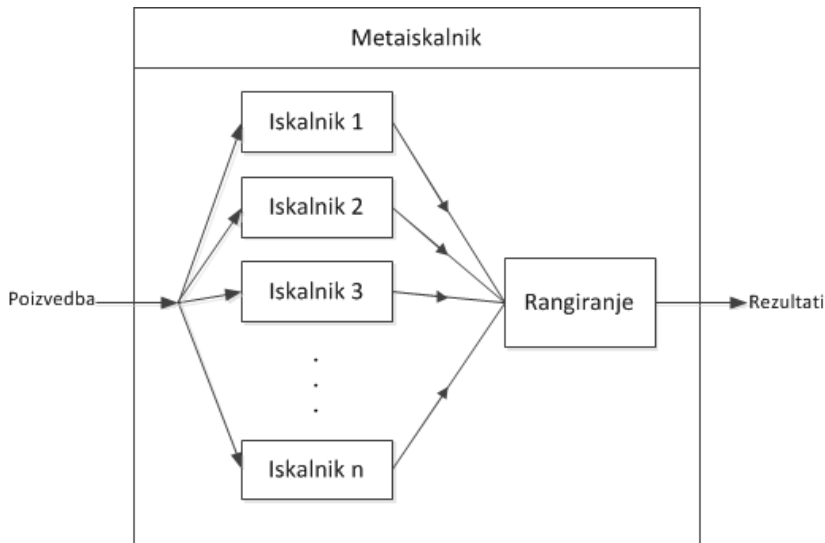


Susan Dumais
2009

Iskalnik in metaiskalnik



Iskalnik in metaiskalnik

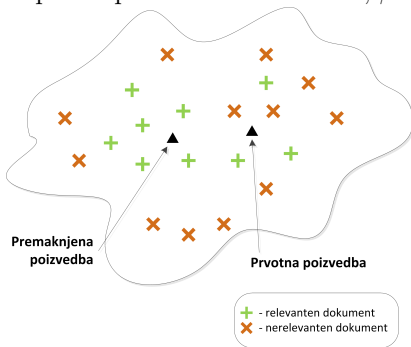


Rocchio algoritem

D_r , D_{nr} iz majhnega dela rezultatov, ki jih je uporabnik označi.

$$\vec{q}_m = \alpha \vec{q}_0 + \beta \frac{1}{|D_r|} \sum_{\vec{d}_j \in D_r} \vec{d}_j - \gamma \frac{1}{|D_{nr}|} \sum_{\vec{d}_j \in D_{nr}} \vec{d}_j \quad (1)$$

V praksi pametna izbira $\alpha = 1$, $\beta = 0,75$ in $\gamma = 0,15$.



Jezikovni model

$q = q_1 q_2 \dots q_m$ poizvedba, d_j dokument. Bayesovo pravilo:

$$Pr(d_j|q) = \frac{Pr(q|d_j)Pr(d_j)}{Pr(q)} \quad (2)$$

Poenostavitev: $Pr(d_j|q) = Pr(q|d_j)$. Predpostavimo neodvisnost besed v dokumentih. f_{iq} - število pojavitev besede t_i v poizvedbi q . Za vsak d_j :

$$Pr(q = q_1 q_2 \dots q_m | d_j) = \prod_{i=1}^m Pr(q_i | d_j) = \prod_{i=1}^{|V|} Pr(t_i | d_j)^{f_{iq}} \quad (3)$$

$$Pr(t_i | d_j) = \frac{f_{ij}}{|d_j|} \quad (4)$$

Laplaceova ocena ($\lambda = 1$):

$$Pr(t_i | d_j) = \frac{\lambda + f_{ij}}{\lambda|V| + |d_j|} \quad (5)$$

Latentno semantično indeksiranje

m - število možnih besed, n - število dokumentov v sistemu, A matrika $m \times n$. Singularni razcep:

$$A = U\Sigma V^T \quad (6)$$

$U - m \times r$, $\Sigma - r \times r$, $V - n \times r$ Lastnost razcepa: aproksimacija A s $k \leq \text{rang}(A)$ dimenzijami:

$$A_k = U_k \Sigma_k V_k^T \quad (7)$$

Pretvorba poizvedbe:

$$q = U_k \Sigma_k q_k^T \quad (8)$$

$$q_k = q^T U_k \Sigma_k^{-1} \quad (9)$$

Pridobitev podobnosti: primerjava q_k z vrsticami (dokumenti) V_k

Latentno semantično indeksiranje - primer

V sistemu imamo 6 dokumentov, ki vsebujejo 5 različnih besed.

	d_1	d_2	d_3	d_4	d_5	d_6
$A =$ ladja	1	0	1	0	0	0
čoln	0	1	0	0	0	0
ocean	1	1	0	0	0	0
križarjenje	1	0	0	1	1	0
jadranje	0	0	0	1	0	1

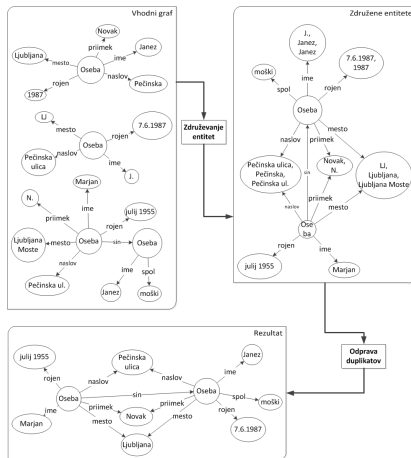
Latentno semantično indeksiranje - primer

		1	2	3	4	5	
$U =$	ladja						
	čoln						
	ocean						
	križarjenje						
	jadranje						
		d_1	d_2	d_3	d_4	d_5	d_6
$V^T =$	1						
	2						
	3						
	4						
	5						

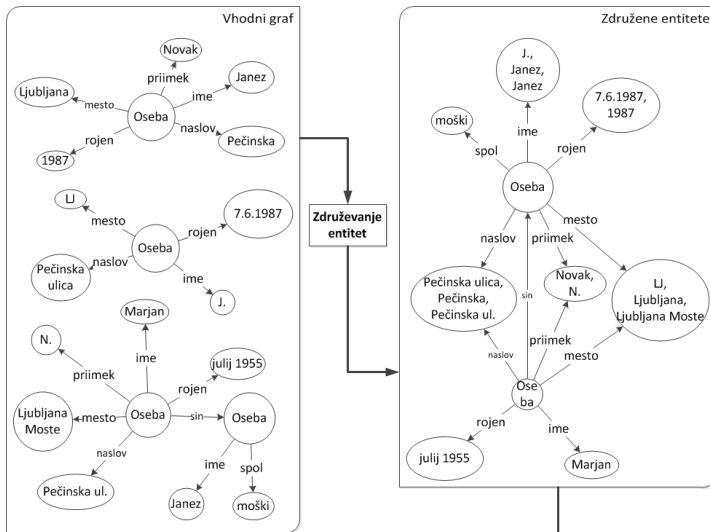
Σ - diagonalna s singularnimi vrednostmi 5×5

Primer v Mathematici.

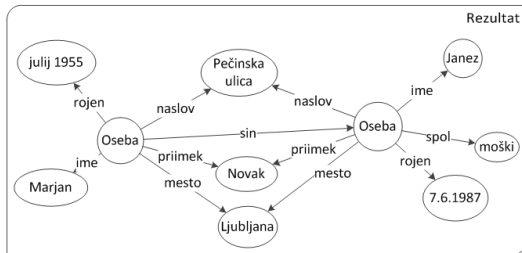
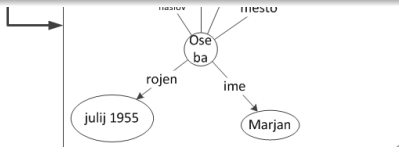
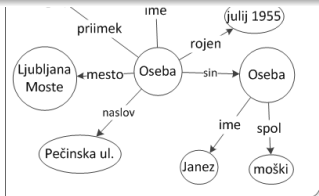
Združevanje entitet in odprava duplikatov



Združevanje entitet in odprava duplikatov



Združevanje entitet in odprava duplikatov



Odprava
duplikatov

PageRank

i - dokument v sistemu, $V(i)$ - množica dokumentov, ki kažejo na i ,
 $I(i)$ - množica dokumentov, na katere kaže i , ocena PageRank:

$$PR(i) = \sum_{j \in V(i)} \frac{PR(j)}{|I(j)|} \quad (10)$$

$$P_{ij} = \begin{cases} \frac{1}{N} & |I(i)| = \emptyset \\ \frac{1}{|I(i)|} & \text{obstaja povezava od } i \text{ do } j \\ 0 & \text{sicer} \end{cases} \quad (11)$$

$$PR(i) = (1 - d) + d \sum_{j \in V(i)} P_{ji} PR(j) \quad (12)$$

PageRank - markovska veriga

$$p_k = P^T p_{k-1} \quad (13)$$

Izrek: Za končno Markovsko verigo s stohastično prehodno matriko P (vsota vsake vrstice je 1, vsak dokument mora imeti vsaj eno izhodno povezavo), ki je aperiodična ($\gcd(\text{dolžin ciklov}) = 1$) in krepko povezana, velja:

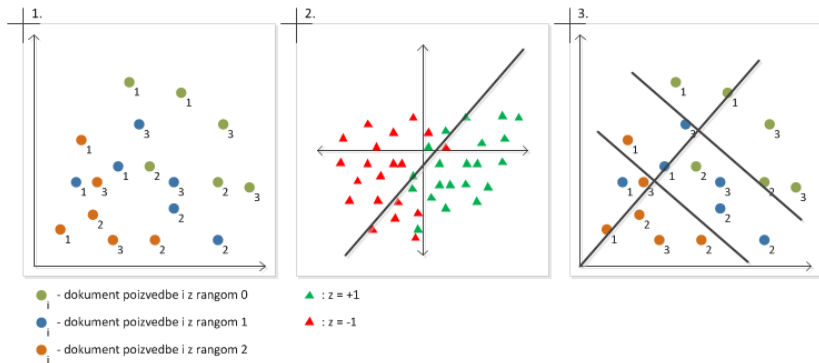
- obstaja ntk. ena stacionarna porazdelitev π
- $f_{i,i} = 1$ in $h_{i,i} = \frac{1}{\pi_i}$
- $\lim_{t \rightarrow \infty} \frac{N(i,t)}{t} = \pi_i$

$$\pi = P^T \pi \quad (14)$$

Ranking SVM

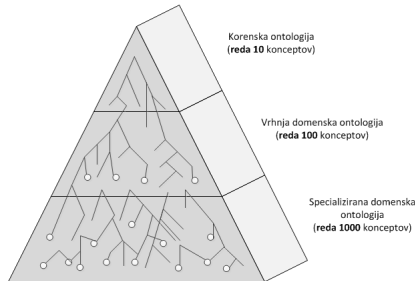
Učni podatki: $(y, q, \vec{x})$, pretvorba v $(\vec{x}_i^{(1)} - \vec{x}_i^{(2)}, z)$ in izvajanje SVM algoritma

$$z = \begin{cases} +1 & \vec{y}_i^{(1)} > \vec{y}_i^{(2)} \\ -1 & \text{sicer} \end{cases} \quad (15)$$



Ontologija

Predstavitev konceptov in razmerij med njimi. Vsebujejo objekte, razrede, attribute, relacije, omejitve, pravila, aksiome. Cilj: predstavitev znanja, semantični splet



Ontologija - poizvedba

```
PREFIX xonto: <http://zitnik.si/druzbenOntologija/0.1/>
SELECT ?ime ?mail
WHERE {
    ?oseba a xonto:Oseba.
    ?oseba xonto:ime      ?ime.
    ?oseba xonto:enaslov  ?mail.
    ?oseba xonto:naslov   "Zgornji Duplek".
}
```

Evaluacija

TF-IDF:

$$tf_{ij} = \frac{f_{ij}}{\max\{f_{1j}, f_{2j}, \dots, f_{|V|j}\}} \quad (16)$$

$$idf_i = \log \frac{n}{df_i} \quad (17)$$

$$w_{ij} = tfidf_{ij} = tf_{ij} \times idf_i \quad (18)$$

		Dejanske vrednosti		
		+	−	
Napovedane vrednosti	+	TP	FP	Napovedani pozitivni
	−	FN	TN	Napovedani negativni
		Pozitivni	Negativni	Vsi

$$CA = \frac{TP + TN}{TP + FP + TN + FN} \quad (19)$$

Evaluacija

$$Natančnost = \frac{\text{število vrnjenih relevantnih}}{\text{število vrnjenih}} = \frac{TP}{TP + FP} \quad (20)$$

$$Prikljic = \frac{\text{število vrnjenih relevantnih}}{\text{število relevantnih v sistemu}} = \frac{TP}{TP + FN} \quad (21)$$

$$E_{\alpha} = 1 - \frac{1}{\frac{\alpha}{Natančnost} + \frac{1-\alpha}{Prikljic}} \quad (22)$$

$$F_{\beta} = (1 + \beta^2) \frac{Natančnost \cdot Prikljic}{\beta^2 Natančnost + Prikljic} \quad (23)$$

$$JaccKoeff(s_i, s_j) = \frac{|Nbr(s_i) \cap Nbr(s_j)|}{|Nbr(s_i) \cup Nbr(s_j)|} \quad (24)$$

Evalvacija

$$MAP(Q) = \frac{1}{|Q|} \sum_{j=1}^{|Q|} \frac{1}{m_j} \sum_{k=1}^{m_j} Natančnost(R_{jk}) \quad (25)$$

$$DCG_p = rel_1 + \sum_{i=2}^p \frac{rel_i}{\log_2 i} \quad (26)$$

$$NDCG_p = \frac{DCG_p}{IdealniDCG_p} \quad (27)$$