Relevance feedback i proširivanje upita

Dragan Ivanović
dragan.ivanovic@uns.ac.rs

Katedra za informatiku, Fakultet tehničkih nauka, Novi Sad

2015.

Relevance feedback: osnovna ideja

- Korisnik postavlja (kratak, jednostavan) upit

Relevance feedback: osnovna ideja

- Korisnik postavlja (kratak, jednostavan) upit
- Pretraživač vraća skup dokumenata

Relevance feedback: osnovna ideja

- Korisnik postavlja (kratak, jednostavan) upit
- Pretraživač vraća skup dokumenata
- Korisnik označava neke dokumente kao relevantne, a neke kao nerelevantne

Relevance feedback: osnovna ideja

- Korisnik postavlja (kratak, jednostavan) upit
- Pretraživač vraća skup dokumenata
- Korisnik označava neke dokumente kao relevantne, a neke kao nerelevantne
- Pretraživač izračunava novu reprezentaciju informacione potrebe, bolju od inicijalnog upita

Relevance feedback: osnovna ideja

- Korisnik postavlja (kratak, jednostavan) upit
- Pretraživač vraća skup dokumenata
- Korisnik označava neke dokumente kao relevantne, a neke kao nerelevantne
- Pretraživač izračunava novu reprezentaciju informacione potrebe, bolju od inicijalnog upita
- Pretraživač sam postavlja novi upit i vraća rezultate korisniku

Relevance feedback: osnovna ideja

- Korisnik postavlja (kratak, jednostavan) upit
- Pretraživač vraća skup dokumenata
- Korisnik označava neke dokumente kao relevantne, a neke kao nerelevantne
- Pretraživač izračunava novu reprezentaciju informacione potrebe, bolju od inicijalnog upita
- Pretraživač sam postavlja novi upit i vraća rezultate korisniku
- Novi rezultati imaju (nadamo se) bolji povrat

Relevance feedback: osnovna ideja

- Korisnik postavlja (kratak, jednostavan) upit
- Pretraživač vraća skup dokumenata
- Korisnik označava neke dokumente kao relevantne, a neke kao nerelevantne
- Pretraživač izračunava novu reprezentaciju informacione potrebe, bolju od inicijalnog upita
- Pretraživač sam postavlja novi upit i vraća rezultate korisniku
- Novi rezultati imaju (nadamo se) bolji povrat
- Ovo može da se radi iterativno

Relevance feedback: osnovna ideja

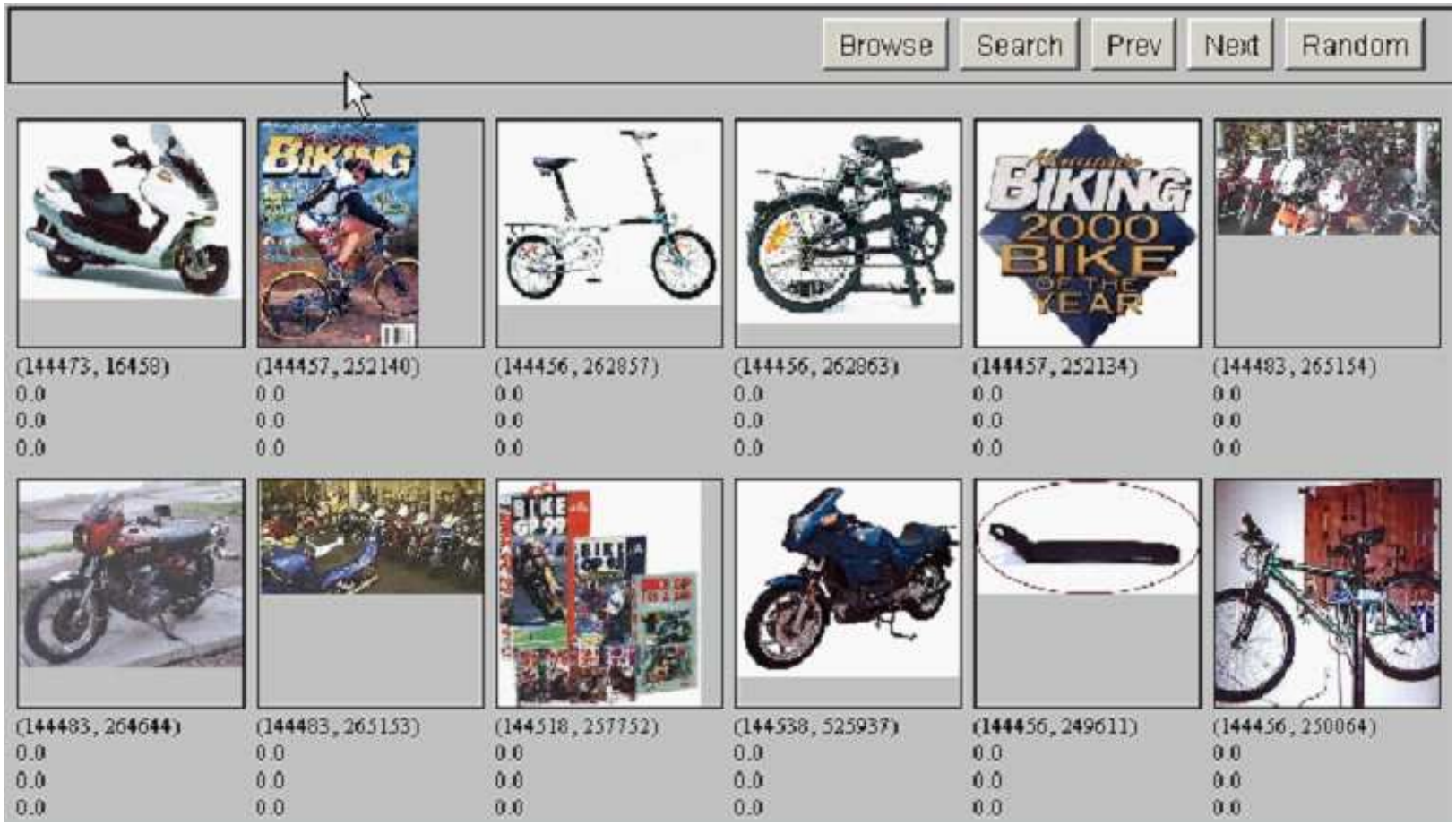
- Korisnik postavlja (kratak, jednostavan) upit
- Pretraživač vraća skup dokumenata
- Korisnik označava neke dokumente kao relevantne, a neke kao nerelevantne
- Pretraživač izračunava novu reprezentaciju informacione potrebe, bolju od inicijalnog upita
- Pretraživač sam postavlja novi upit i vraća rezultate korisniku
- Novi rezultati imaju (nadamo se) bolji povrat
- Ovo može da se radi iterativno
- Obično pretraživanje bez RF zvaćemo *ad hoc* pretraživanje

Relevance feedback: osnovna ideja

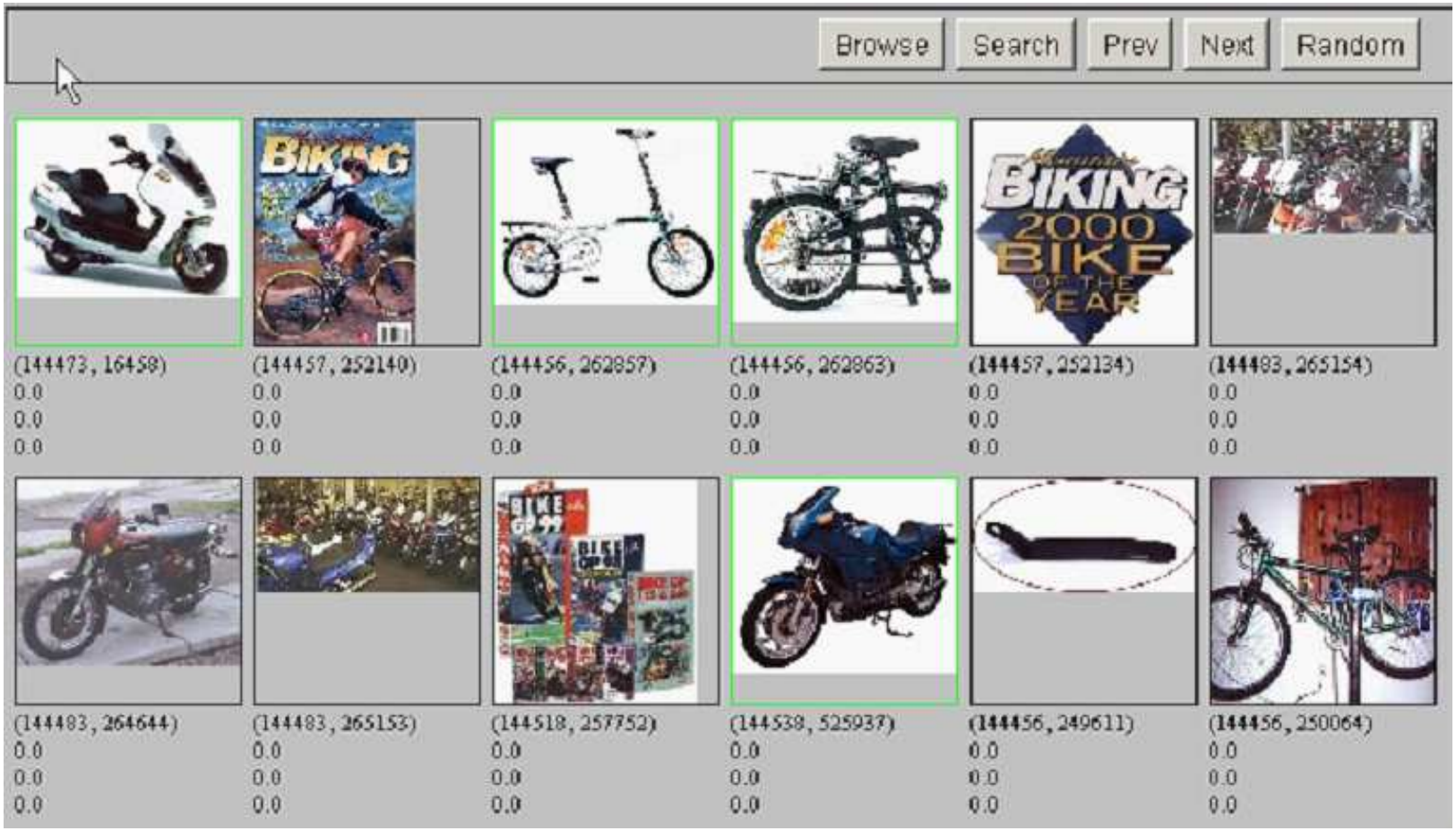
- Korisnik postavlja (kratak, jednostavan) upit
- Pretraživač vraća skup dokumenata
- Korisnik označava neke dokumente kao relevantne, a neke kao nerelevantne
- Pretraživač izračunava novu reprezentaciju informacione potrebe, bolju od inicijalnog upita
- Pretraživač sam postavlja novi upit i vraća rezultate korisniku
- Novi rezultati imaju (nadamo se) bolji povrat
- Ovo može da se radi iterativno
- Obično pretraživanje bez RF zvaćemo *ad hoc* pretraživanje
- Pogledaćemo tri RF primera koji ilustruju različite aspekte procesa

Relevance Feedback: primer

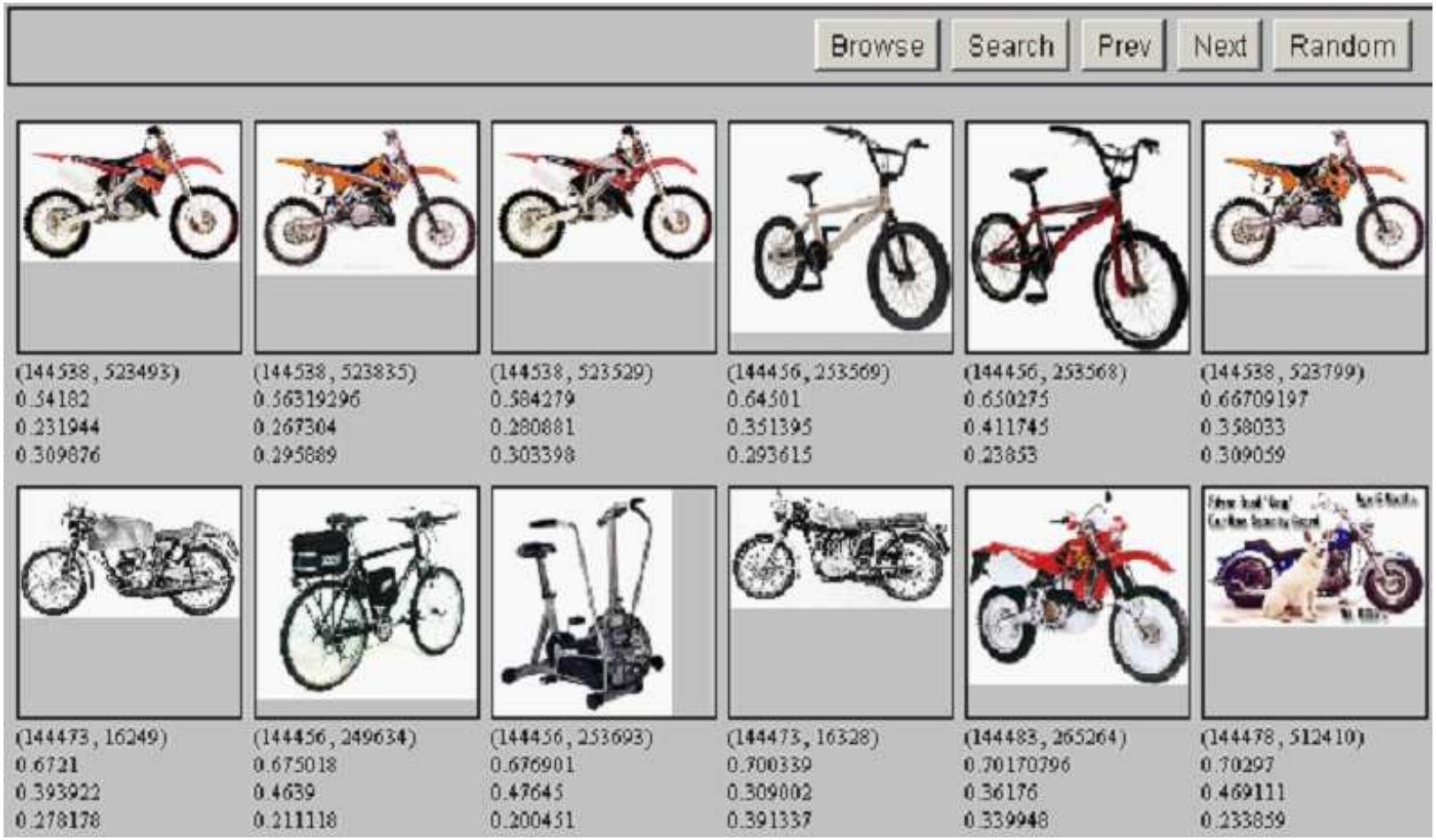
Rezultati za početni upit



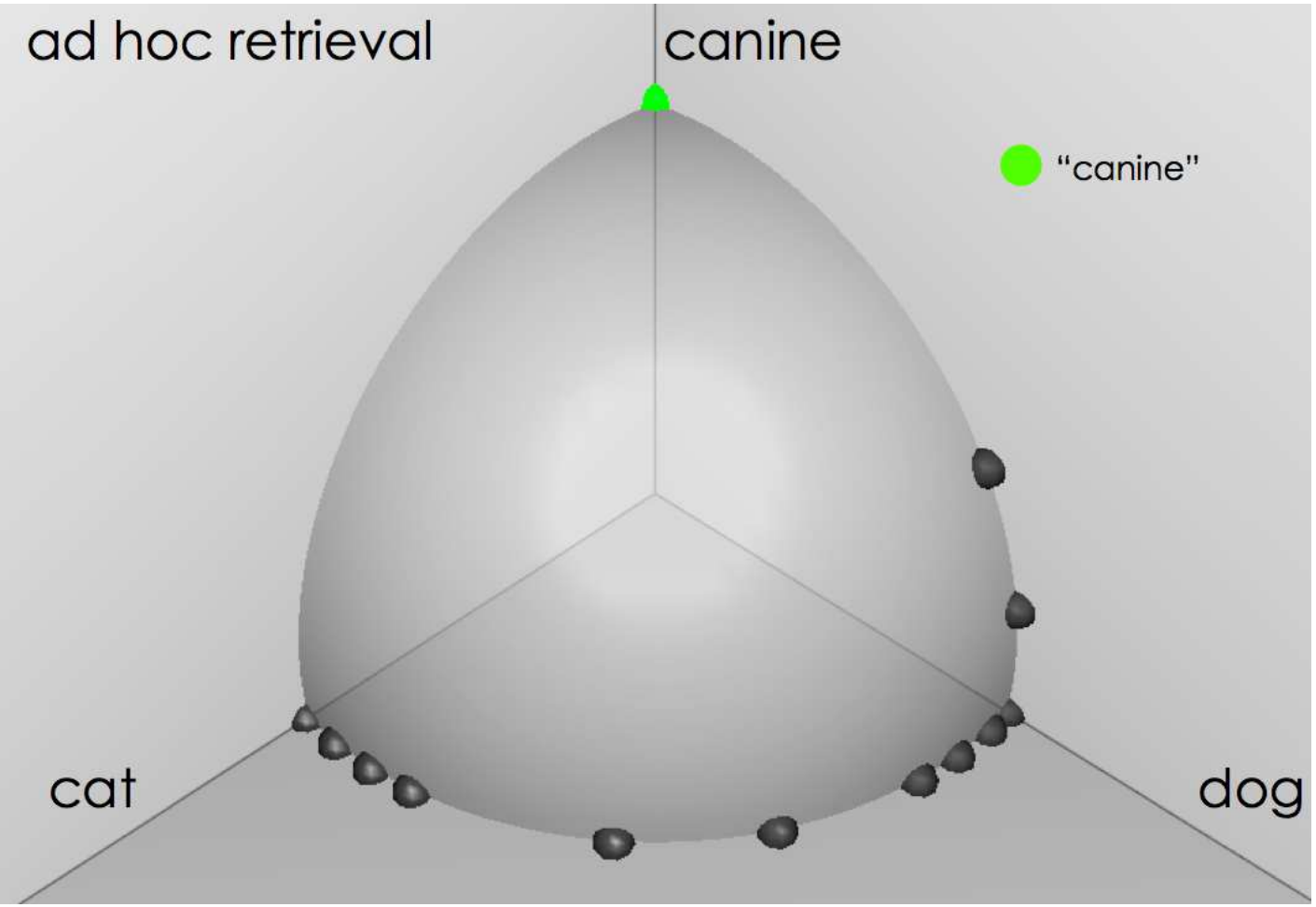
Korisnik označava šta je relevantno



Rezultati posle RF ciklusa

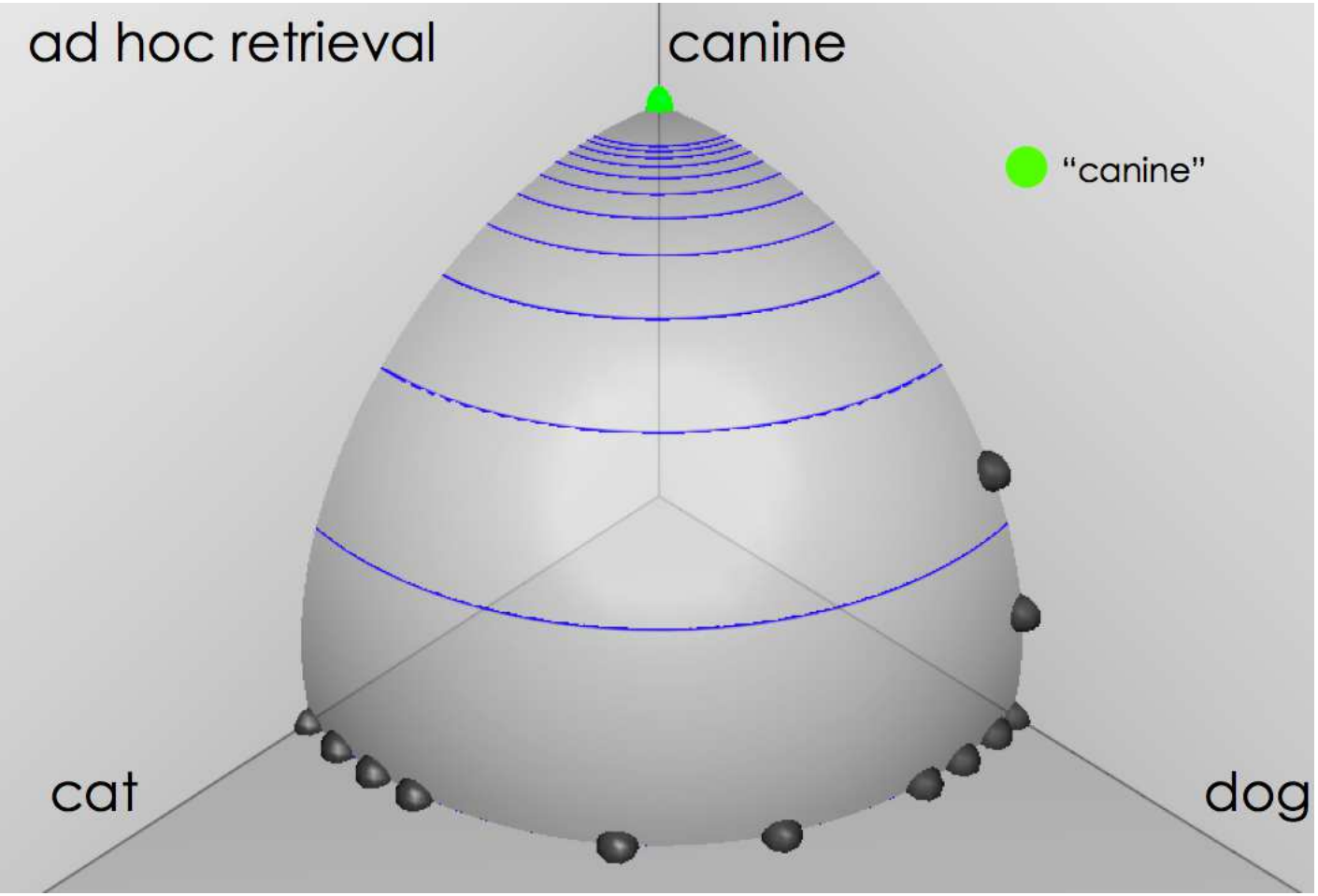


Ad hoc pretraživanje za upit „canine“ (1)



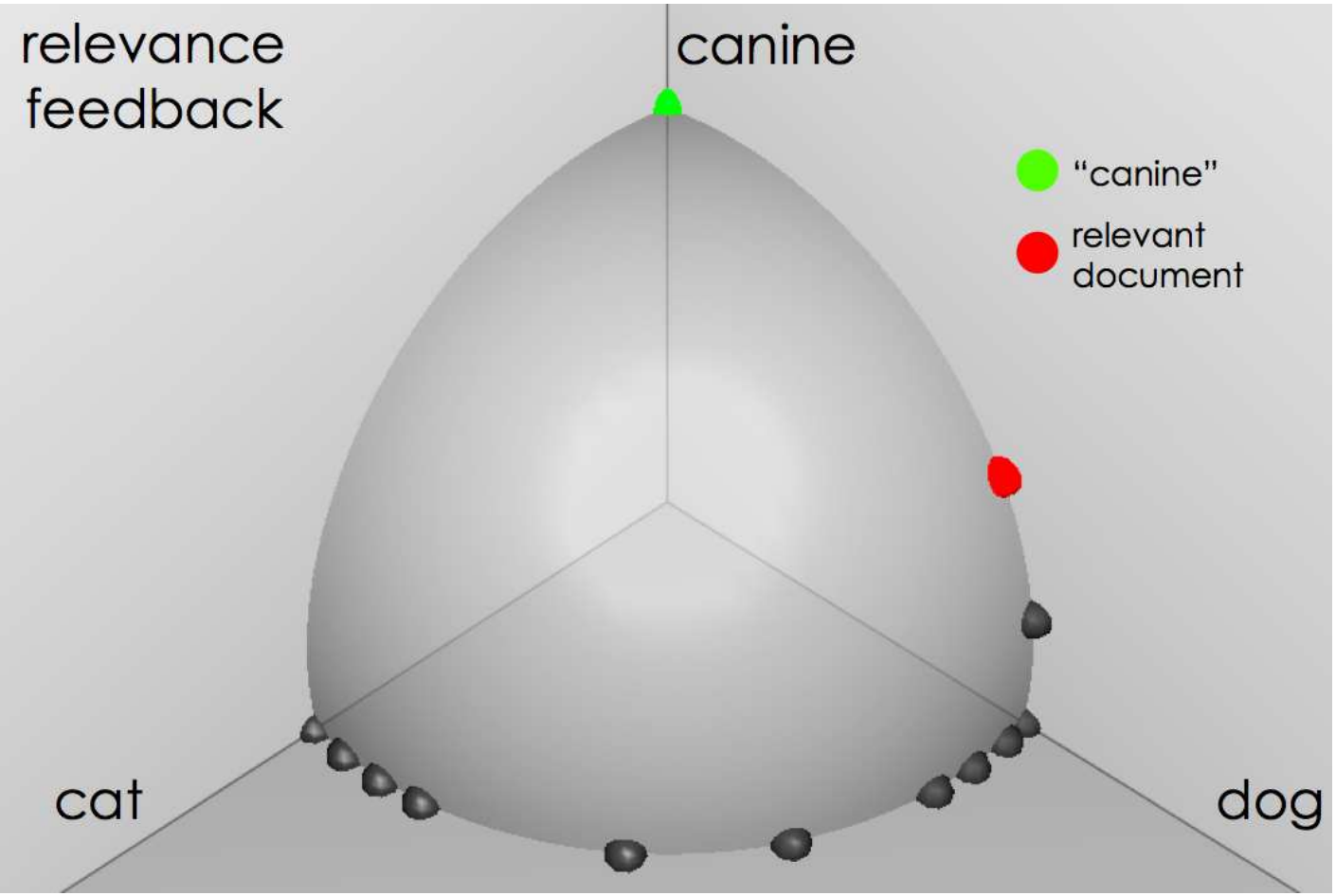
izvor:
Fernando Díaz

Ad hoc pretraživanje za upit „canine“ (2)



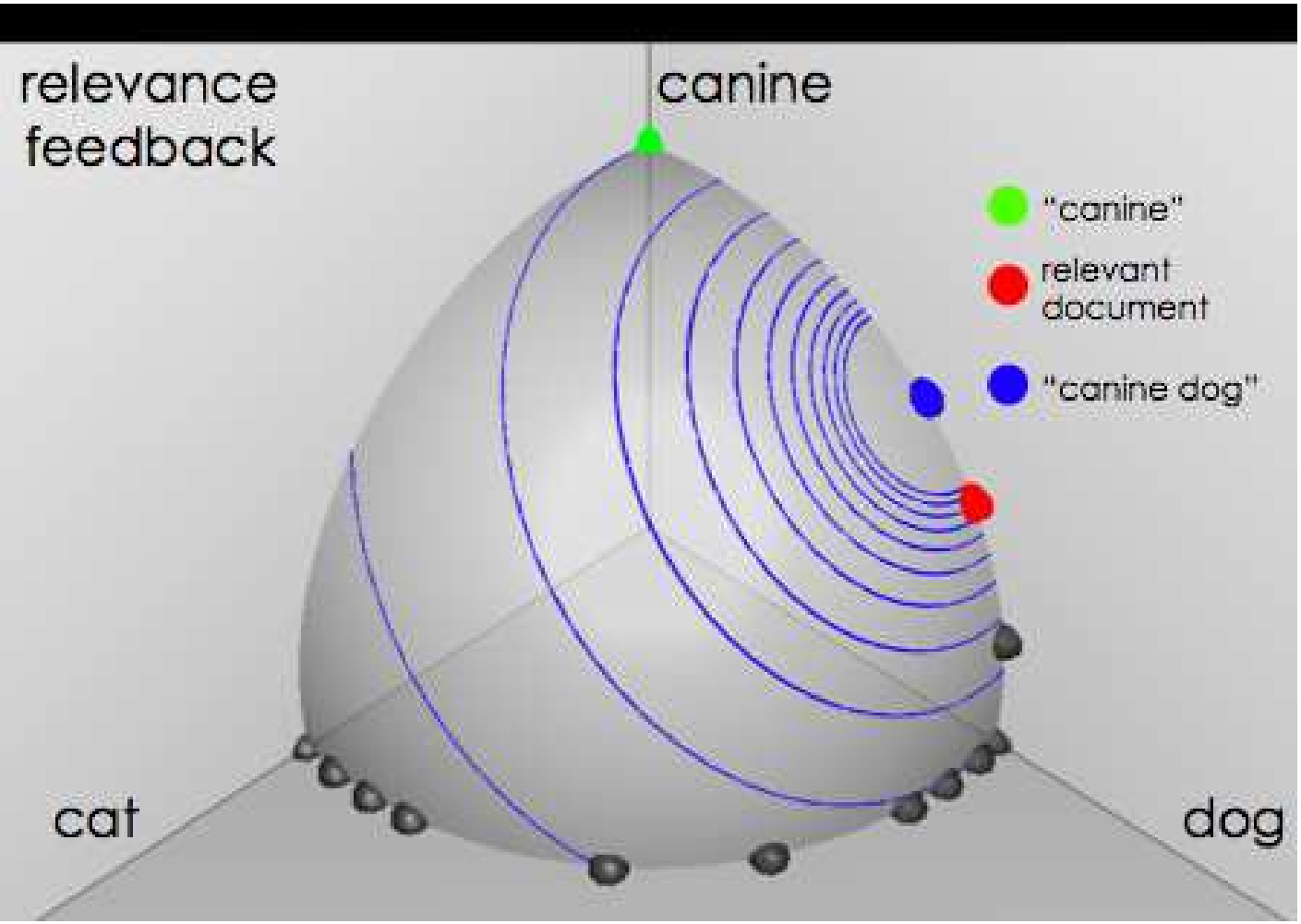
izvor:
Fernando Díaz

RF ciklus: označi relevantne



izvor:
Fernando Díaz

Rezultati posle RF ciklusa



izvor:
Fernando Díaz

Rezultati za početni upit

Početni upit: New space satellite applications

Rezultati za početni upit

Početni upit: New space satellite applications

Rezultati za početni upit:

1. 0.539, 08/13/91, NASA Hasn't Scrapped Imaging Spectrometer
2. 0.533, 07/09/91, NASA Scratches Environment Gear From Satellite Plan
3. 0.528, 04/04/90, Science Panel Backs NASA Satellite Plan, But Urges Launches of Smaller Probes
4. 0.526, 09/09/91, A NASA Satellite Project Accomplishes Incredible Feat: Staying Within Budget
5. 0.525, 07/24/90, Scientist Who Exposed Global Warming Proposes Satellites for Climate Research
6. 0.524, 08/22/90, Report Provides Support for the Critics Of Using Big Satellites to Study Climate
7. 0.516, 04/13/87, Arianespace Receives Satellite Launch Pact From Telesat Canada
8. 0.509, 12/02/87, Telecommunications Tale of Two Companies

Rezultati za početni upit

Početni upit: New space satellite applications

Rezultati za početni upit:

\+ 1. 0.539, 08/13/91, NASA Hasn't Scrapped Imaging Spectrometer

\+ 2. 0.533, 07/09/91, NASA Scratches Environment Gear From Satellite Plan

3. 0.528, 04/04/90, Science Panel Backs NASA Satellite Plan, But Urges Launches of Smaller Probes

4. 0.526, 09/09/91, A NASA Satellite Project Accomplishes Incredible Feat: Staying Within Budget

5. 0.525, 07/24/90, Scientist Who Exposed Global Warming Proposes Satellites for Climate Research

6. 0.524, 08/22/90, Report Provides Support for the Critics Of Using Big Satellites to Study Climate

7. 0.516, 04/13/87, Arianespace Receives Satellite Launch Pact From Telesat Canada

\+ 8. 0.509, 12/02/87, Telecommunications Tale of Two Companies

Korisnik označava relevantne dokumente sa „+“

Prošireni upit posle RF

2.074	new	15.106	space
30.816	satellite	5.660	application
5.991	nasa	5.196	eos
4.196	launch	3.972	aster
3.516	instrument	3.446	arianespace
3.004	bundespost	2.806	ss
2.790	rocket	2.053	scientist
2.003	broadcast	1.172	earth
0.836	oil	0.646	measure

Rezultati za prošireni upit

* 1. 0.513, 07/09/91, NASA Scratches Environment Gear From Satellite Plan
* 2. 0.500, 08/13/91, NASA Hasn't Scrapped Imaging Spectrometer

3. 0.493, 08/07/89, When the Pentagon Launches a Secret Satellite, Space Sleuths Do Some Spy Work of Their Own

4. 0.493, 07/31/89, NASA Uses 'Warm' Superconductors For Fast Circuit

* 5. 0.492, 12/02/87, Telecommunications Tale of Two Companies

6. 0.491, 07/09/91, Soviets May Adapt Parts of SS-20 Missile For Commercial Use

7. 0.490, 07/12/88, Gaping Gap: Pentagon Lags in Race To Match the Soviets In Rocket Launchers

8. 0.490, 06/14/90, Rescue of Satellite By Space Agency To Cost $90 Million

Osnovni koncept za RF: centroid

- Centroid je centar mase za skup tačaka

Osnovni koncept za RF: centroid

- Centroid je centar mase za skup tačaka
- Dokumenti su predstavljeni kao tačke u visoko-dimenzionalnom prostoru

Osnovni koncept za RF: centroid

- Centroid je centar mase za skup tačaka
- Dokumenti su predstavljeni kao tačke u visoko-dimenzionalnom prostoru
- Možemo izračunati centroid za skup dokumenata

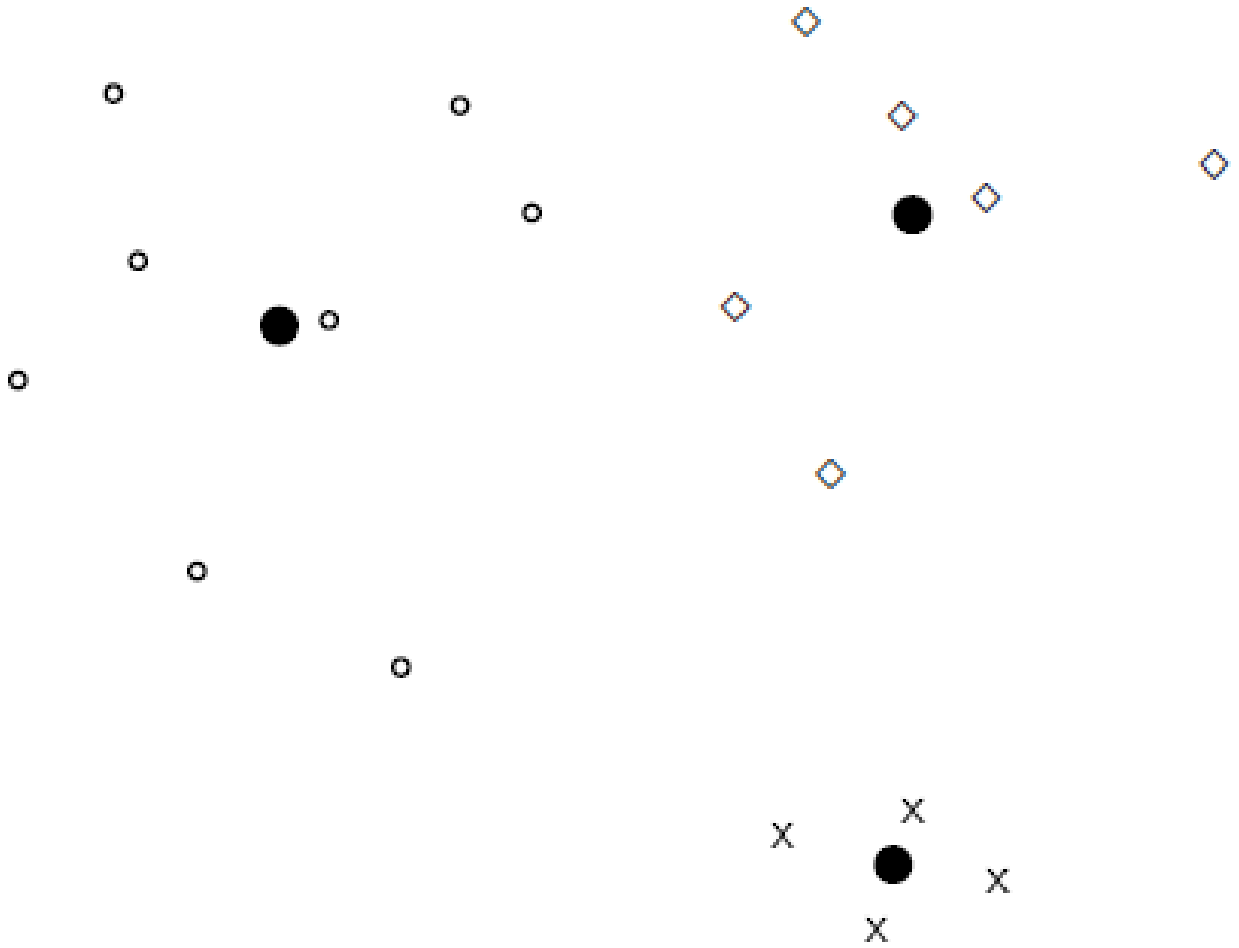
Osnovni koncept za RF: centroid

- Centroid je centar mase za skup tačaka
- Dokumenti su predstavljeni kao tačke u visoko-dimenzionalnom prostoru
- Možemo izračunati centroid za skup dokumenata
- Definicija:

$$\vec{\mu}(D) = \frac{1}{|D|} \sum_{d \in D} \vec{v}(d)$$

gde je D skup dokumenata i $\vec{v}(d) = \vec{d}$ je vektor kojim reprezentujemo dokument d

Centroid: primeri



Rocchio-ov algoritam

- Rocchio-ov algoritam implementira RF u vektorskom modelu

Rocchio-ov algoritam

- Rocchio-ov algoritam implementira RF u vektorskom modelu
- Rocchio bira upit $\vec{q}_{opt}$ koji maksimizuje

$$\vec{q}_{opt} = \max_{\vec{q}}[\text{sim}(\vec{q}, D_r) - \text{sim}(\vec{q}, D_{nr})]$$

Rocchio-ov algoritam

- Rocchio-ov algoritam implementira RF u vektorskom modelu
- Rocchio bira upit $\vec{q}_{opt}$ koji maksimizuje

$$\vec{q}_{opt} = \max_{\vec{q}}[\text{sim}(\vec{q}, D_r) - \text{sim}(\vec{q}, D_{nr})]$$

- Ideja je da se napravi maksimalno razdvajanje relevantnih i nerelevantnih dokumenata

Rocchio-ov algoritam

- Rocchio-ov algoritam implementira RF u vektorskom modelu
- Rocchio bira upit $\vec{q}_{opt}$ koji maksimizuje

$$\vec{q}_{opt} = \max_{\vec{q}}[\text{sim}(\vec{q}, D_r) - \text{sim}(\vec{q}, D_{nr})]$$

- Ideja je da se napravi maksimalno razdvajanje relevantnih i nerelevantnih dokumenata
- Optimalni vektor upita je

$$\vec{q}_{opt} = \frac{1}{|D_r|}\sum_{\vec{d}_j \in D_r}\vec{d}_j + [\frac{1}{|D_r|}\sum_{\vec{d}_j \in D_r}\vec{d}_j - \frac{1}{|D_{nr}|}\sum_{\vec{d}_j \in D_{nr}}\vec{d}_j]$$

D_r: skup relevantnih dokumenata; D_{nr}: skup nerelevantnih dokumenata

Rocchio-ov algoritam

- Optimalni vektor upita je

$$\vec{q}_{opt} = \frac{1}{|D_r|} \sum_{\vec{d}_j \in D_r} \vec{d}_j + [\frac{1}{|D_r|} \sum_{\vec{d}_j \in D_r} \vec{d}_j - \frac{1}{|D_{nr}|} \sum_{\vec{d}_j \in D_{nr}} \vec{d}_j]$$

Rocchio-ov algoritam

- Optimalni vektor upita je

$$\vec{q}_{opt} = \frac{1}{|D_r|} \sum_{\vec{d}_j \in D_r} \vec{d}_j + [\frac{1}{|D_r|} \sum_{\vec{d}_j \in D_r} \vec{d}_j - \frac{1}{|D_{nr}|} \sum_{\vec{d}_j \in D_{nr}} \vec{d}_j]$$

- q-opt = centroid-rel + (centroid-rel - centroid-nonrel)

Rocchio-ov algoritam

- Optimalni vektor upita je

$$\vec{q}_{opt} = \frac{1}{|D_r|} \sum_{\vec{d}_j \in D_r} \vec{d}_j + [\frac{1}{|D_r|} \sum_{\vec{d}_j \in D_r} \vec{d}_j - \frac{1}{|D_{nr}|} \sum_{\vec{d}_j \in D_{nr}} \vec{d}_j]$$

- q-opt = centroid-rel + (centroid-rel - centroid-nonrel)
- Pomeramo centroid relevantnih dokumenata za razliku dva centroida

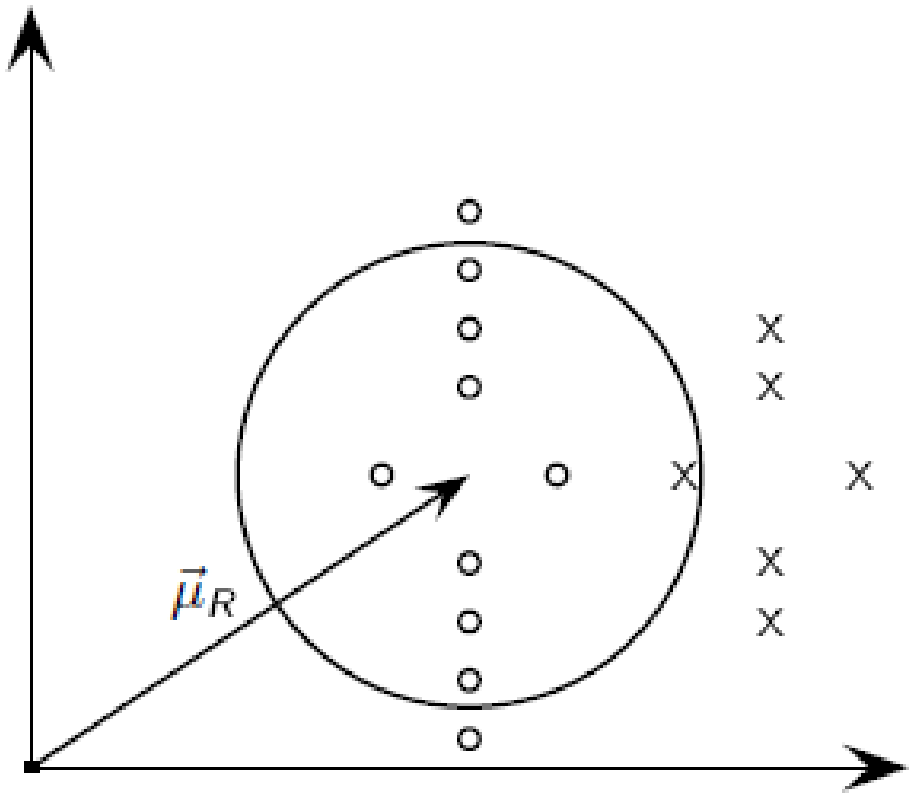
Rocchio-ov algoritam

- Optimalni vektor upita je

$$\vec{q}_{opt} = \frac{1}{|D_r|} \sum_{\vec{d}_j \in D_r} \vec{d}_j + [\frac{1}{|D_r|} \sum_{\vec{d}_j \in D_r} \vec{d}_j - \frac{1}{|D_{nr}|} \sum_{\vec{d}_j \in D_{nr}} \vec{d}_j]$$

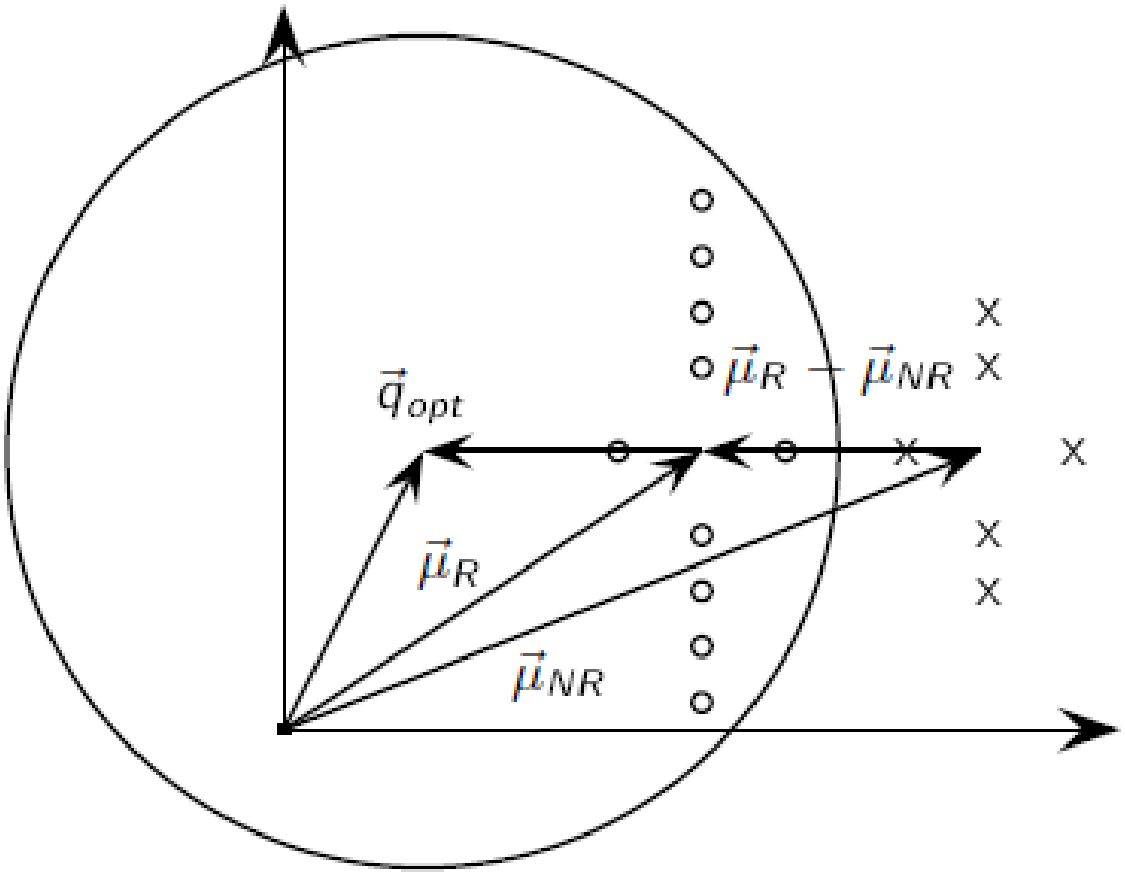
- q-opt = centroid-rel + (centroid-rel - centroid-nonrel)
- Pomeramo centroid relevantnih dokumenata za razliku dva centroida
- Pretpostavili smo da $|\vec{\mu}_r| = |\vec{\mu}_{nr}| = 1$ za ovaj slučaj.

Rocchio ilustracija



$\vec{\mu}_R$ ne razdvaja relevantne i nerelevantne

Rocchio ilustracija



$\vec{q}_{opt}$ odlično razdvaja relevantne i nerelevantne

Rocchio 1971 algoritam (SMART sistem)

- Problem je što mi obično ne znamo sve relevantne i nerelevantne dokumente, i ne znamo da nađemo njihov centroid

Rocchio 1971 algoritam (SMART sistem)

- Problem je što mi obično ne znamo sve relevantne i nerelevantne dokumente, i ne znamo da nađemo njihov centroid
- Ono što se koristi u praksi

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

q_m: modifikovani vektor upita q_0: originalni vektor upita; D_r i D_{nr}: skup poznatih relevantnih odnosno nerelevantnih α, β i γ: težine

Rocchio 1971 algoritam (SMART sistem)

- Problem je što mi obično ne znamo sve relevantne i nerelevantne dokumente, i ne znamo da nađemo njihov centroid
- Ono što se koristi u praksi

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

  q_m: modifikovani vektor upita q_0: originalni vektor upita; D_r i D_{nr}: skup poznatih relevantnih odnosno nerelevantnih α, β i γ: težine
- Novi upit se pomera prema relevantnim dokumentima i udaljava od nerelevantnih

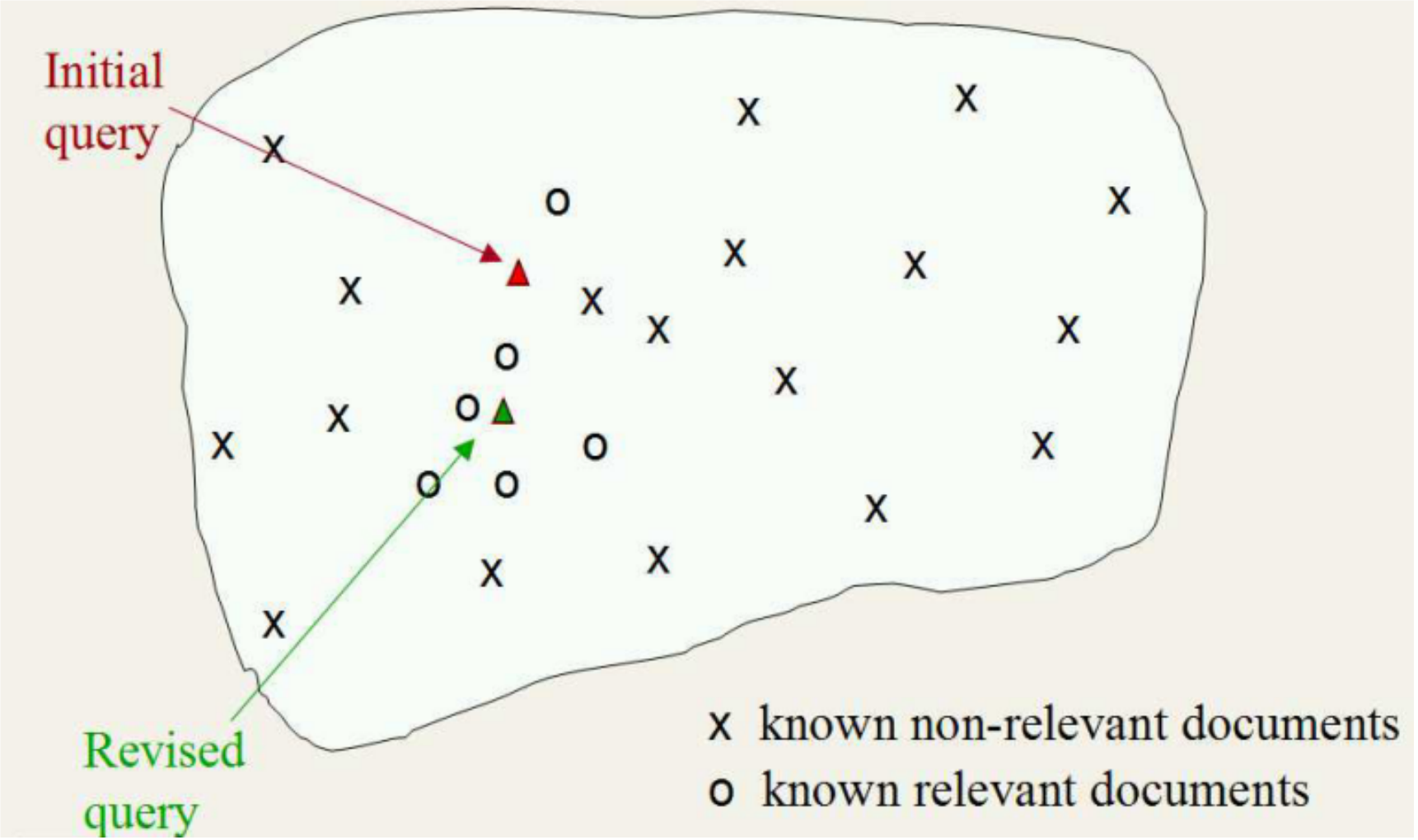
Rocchio 1971 algoritam (SMART sistem)

- Problem je što mi obično ne znamo sve relevantne i nerelevantne dokumente, i ne znamo da nađemo njihov centroid
- Ono što se koristi u praksi

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

  q_m: modifikovani vektor upita q_0: originalni vektor upita; D_r i D_{nr}: skup poznatih relevantnih odnosno nerelevantnih α, β i γ: težine
- Novi upit se pomera prema relevantnim dokumentima i udaljava od nerelevantnih
- Kompromis α vs. β/γ: Ako imamo mnogo ocenjenih dokumenata, treba nam veći β/γ

Rocchio RF ilustracija



Pozitivni i negativni RF

- Pozitivni odgovori su vredniji nego negativni

Pozitivni i negativni RF

- Pozitivni odgovori su vredniji nego negativni
- Zašto?

Pozitivni i negativni RF

- Pozitivni odgovori su vredniji nego negativni
- Zašto?
- Na primer, neka je $\beta = 0.75$, $\gamma = 0.25$ da se dobije veća težina za pozitivni odgovor

Pozitivni i negativni RF

- Pozitivni odgovori su vredniji nego negativni
- Zašto?
- Na primer, neka je $\beta = 0.75$, $\gamma = 0.25$ da se dobije veća težina za pozitivni odgovor
- Mnogi sistemi dozvoljavaju samo pozitivne odgovore

Relevance feedback: pretpostavke

- Kada RF može da unapredi povrat?

Relevance feedback: pretpostavke

- Kada RF može da unapredi povrat?
- Pretpostavka 1: korisnik zna termove u kolekciji dovoljno dobro za početni upit

Relevance feedback: pretpostavke

- Kada RF može da unapredi povrat?
- Pretpostavka 1: korisnik zna termove u kolekciji dovoljno dobro za početni upit
- Pretpostavka 2: relevantni dokumenti sadrže slične termove (tako da možemo „skakati“ od jednog relevantnog dokumenta na drugi kad unosimo RF)

Kršenje pretpostavke 1

- Kršenje pretpostavke 1: korisnik zna termove u kolekciji dovoljno dobro za početni upit

Kršenje pretpostavke 1

- Kršenje pretpostavke 1: korisnik zna termove u kolekciji dovoljno dobro za početni upit
- Neslaganje korisnikovog rečnika i rečnika kolekcije

Kršenje pretpostavke 1

- Kršenje pretpostavke 1: korisnik zna termove u kolekciji dovoljno dobro za početni upit
- Neslaganje korisnikovog rečnika i rečnika kolekcije
- Primer: kosmonaut / astronaut

Kršenje pretpostavke 2

- Kršenje pretpostavke 2: relevantni dokumenti nisu slični

Kršenje pretpostavke 2

- Kršenje pretpostavke 2: relevantni dokumenti nisu slični
- Primer upita: contradictory government policies

Kršenje pretpostavke 2

- Kršenje pretpostavke 2: relevantni dokumenti nisu slični
- Primer upita: contradictory government policies
- Zašto RF neće značajno povećati povrat za ovaj upit?

Kršenje pretpostavke 2

- Kršenje pretpostavke 2: relevantni dokumenti nisu slični
- Primer upita: contradictory government policies
- Zašto RF neće značajno povećati povrat za ovaj upit?
- Nekoliko nevezanih „prototipova“

Kršenje pretpostavke 2

- Kršenje pretpostavke 2: relevantni dokumenti nisu slični
- Primer upita: contradictory government policies
- Zašto RF neće značajno povećati povrat za ovaj upit?
- Nekoliko nevezanih „prototipova“
  - Subsidies for tobacco farmers vs. anti-smoking campaigns

Kršenje pretpostavke 2

- Kršenje pretpostavke 2: relevantni dokumenti nisu slični
- Primer upita: contradictory government policies
- Zašto RF neće značajno povećati povrat za ovaj upit?
- Nekoliko nevezanih „prototipova“
  - Subsidies for tobacco farmers vs. anti-smoking campaigns
  - Aid for developing countries vs. high tariffs on imports from developing countries

Kršenje pretpostavke 2

- Kršenje pretpostavke 2: relevantni dokumenti nisu slični
- Primer upita: contradictory government policies
- Zašto RF neće značajno povećati povrat za ovaj upit?
- Nekoliko nevezanih „prototipova“
  - Subsidies for tobacco farmers vs. anti-smoking campaigns
  - Aid for developing countries vs. high tariffs on imports from developing countries
- RF na „tobacco“ dokumente neće pomoći u pronalaženju dokumenata o „developing countries“

Relevance feedback: vrednovanje

- Uzmimo jednu od mera za performanse IR sistema, recimo preciznost za najboljih 10 dokumenata: $P@10$

Relevance feedback: vrednovanje

- Uzmimo jednu od mera za performanse IR sistema, recimo preciznost za najboljih 10 dokumenata: $P@10$
- Izračunamo $P@10$ za originalni upit q_0

Relevance feedback: vrednovanje

- Uzmimo jednu od mera za performanse IR sistema, recimo preciznost za najboljih 10 dokumenata: $P@10$
- Izračunamo $P@10$ za originalni upit q_0
- Izračunamo $P@10$ za RF-modifikovani upit q_1

Relevance feedback: vrednovanje

- Uzmimo jednu od mera za performanse IR sistema, recimo preciznost za najboljih 10 dokumenata: $P@10$
- Izračunamo $P@10$ za originalni upit q_0
- Izračunamo $P@10$ za RF-modifikovani upit q_1
- U većini slučajeva: q_1 je drastično bolje od q_0!

Relevance feedback: vrednovanje

- Uzmimo jednu od mera za performanse IR sistema, recimo preciznost za najboljih 10 dokumenata: $P@10$
- Izračunamo $P@10$ za originalni upit q_0
- Izračunamo $P@10$ za RF-modifikovani upit q_1
- U većini slučajeva: q_1 je drastično bolje od q_0!
- Da li je ovo fer poređenje?

Relevance feedback: vrednovanje

- Fer poređenje mora biti na ostatku kolekcije: dokumentima koje korisnik još nije ocenio

Relevance feedback: vrednovanje

- Fer poređenje mora biti na ostatku kolekcije: dokumentima koje korisnik još nije ocenio
- Studije pokazuju da je RF uspešan ako se vrednuje na ovakav način

Relevance feedback: vrednovanje

- Fer poređenje mora biti na ostatku kolekcije: dokumentima koje korisnik još nije ocenio
- Studije pokazuju da je RF uspešan ako se vrednuje na ovakav način
- Empirijski, jedan RF ciklus je često vrlo uspešan; drugi je marginalno koristan

Relevance feedback: vrednovanje

- Pravo vrednovanje koristi mora uključiti druge metode koje troše istu količinu vremena.

Relevance feedback: vrednovanje

- Pravo vrednovanje koristi mora uključiti druge metode koje troše istu količinu vremena.
- Alternativa za RF: korisnik revidira i ponovo pošalje upit

Relevance feedback: vrednovanje

- Pravo vrednovanje koristi mora uključiti druge metode koje troše istu količinu vremena.
- Alternativa za RF: korisnik revidira i ponovo pošalje upit
- Korisnicima može da se više sviđa revizija upita nego ocenjivanje dokumenata

Relevance feedback: vrednovanje

- Pravo vrednovanje koristi mora uključiti druge metode koje troše istu količinu vremena.
- Alternativa za RF: korisnik revidira i ponovo pošalje upit
- Korisnicima može da se više sviđa revizija upita nego ocenjivanje dokumenata
- Nema jasnog dokaza da je RF „najbolje korišćenje“ korisnikovog vremena

Da li web pretraživači koriste RF?

Google: „similar pages“

Relevance feedback: problemi

- Relevance feedback je skup

Relevance feedback: problemi

- Relevance feedback je skup
  - RF kreira dugačke modifikovane upite

Relevance feedback: problemi

- Relevance feedback je skup
  - RF kreira dugačke modifikovane upite
  - dugački upiti su skupi za obradu

Relevance feedback: problemi

- Relevance feedback je skup
  - RF kreira dugačke modifikovane upite
  - dugački upiti su skupi za obradu
- Korisnici oklevaju da daju eksplicitne odgovore

Relevance feedback: problemi

- Relevance feedback je skup
  - RF kreira dugačke modifikovane upite
  - dugački upiti su skupi za obradu
- Korisnici oklevaju da daju eksplicitne odgovore
- Često je teško razumeti zašto je neki dokument pronađen nakon što je primenjen RF

Druge upotrebe za RF

- Održavanje trajnih upita

Druge upotrebe za RF

- Održavanje trajnih upita
- Primer: “multicore computer chips”

Druge upotrebe za RF

- Održavanje trajnih upita
- Primer: “multicore computer chips”
- Želim da primim svakog dana listu novinskih članaka objavljenih u prethodnih 24 sata na temu „multicore computer chips“

Druge upotrebe za RF

- Održavanje trajnih upita
- Primer: “multicore computer chips”
- Želim da primim svakog dana listu novinskih članaka objavljenih u prethodnih 24 sata na temu „multicore computer chips“
- RF se može koristiti za rafinaciju ovog trajnog upita tokom vremena

Druge upotrebe za RF

- Održavanje trajnih upita
- Primer: “multicore computer chips”
- Želim da primim svakog dana listu novinskih članaka objavljenih u prethodnih 24 sata na temu „multicore computer chips“
- RF se može koristiti za rafinaciju ovog trajnog upita tokom vremena
- Spam filteri rade sličnu stvar

Druge upotrebe za RF

- Održavanje trajnih upita
- Primer: “multicore computer chips”
- Želim da primim svakog dana listu novinskih članaka objavljenih u prethodnih 24 sata na temu „multicore computer chips“
- RF se može koristiti za rafinaciju ovog trajnog upita tokom vremena
- Spam filteri rade sličnu stvar
- RF je mnogo praktičniji za trajne upite nego za web pretragu

Pseudo-RF

- *Blind relevance feedback*

Pseudo-RF

- *Blind relevance feedback*
- Pseudo-RF automatizuje „ručni“ deo pravog RF ciklusa

Pseudo-RF

- *Blind relevance feedback*
- Pseudo-RF automatizuje „ručni“ deo pravog RF ciklusa
- Pseudo-RF algoritam:

Pseudo-RF

- *Blind relevance feedback*
- Pseudo-RF automatizuje „ručni“ deo pravog RF ciklusa
- Pseudo-RF algoritam:
  - Izračunaj rangiranu listu pogodaka za korisnikov upit

Pseudo-RF

- *Blind relevance feedback*
- Pseudo-RF automatizuje „ručni“ deo pravog RF ciklusa
- Pseudo-RF algoritam:
  - Izračunaj rangiranu listu pogodaka za korisnikov upit
  - Pretpostavi da je najboljih k dokumenata relevantno

Pseudo-RF

- *Blind relevance feedback*
- Pseudo-RF automatizuje „ručni“ deo pravog RF ciklusa
- Pseudo-RF algoritam:
  - Izračunaj rangiranu listu pogodaka za korisnikov upit
  - Pretpostavi da je najboljih k dokumenata relevantno
  - Uradi RF ciklus (npr. Rocchio algoritmom)

Pseudo-RF

- *Blind relevance feedback*
- Pseudo-RF automatizuje „ručni“ deo pravog RF ciklusa
- Pseudo-RF algoritam:
  - Izračunaj rangiranu listu pogodaka za korisnikov upit
  - Pretpostavi da je najboljih k dokumenata relevantno
  - Uradi RF ciklus (npr. Rocchio algoritmom)
- Radi dobro u proseku, ali može da napravi katastrofu za neke upite

Pseudo-RF

- *Blind relevance feedback*
- Pseudo-RF automatizuje „ručni“ deo pravog RF ciklusa
- Pseudo-RF algoritam:
  - Izračunaj rangiranu listu pogodaka za korisnikov upit
  - Pretpostavi da je najboljih k dokumenata relevantno
  - Uradi RF ciklus (npr. Rocchio algoritmom)
- Radi dobro u proseku, ali može da napravi katastrofu za neke upite
- Više iteracija može da izazove *klizanje upita*.

Pseudo-RF

- *Blind relevance feedback*
- Pseudo-RF automatizuje „ručni“ deo pravog RF ciklusa
- Pseudo-RF algoritam:
  - Izračunaj rangiranu listu pogodaka za korisnikov upit
  - Pretpostavi da je najboljih k dokumenata relevantno
  - Uradi RF ciklus (npr. Rocchio algoritmom)
- Radi dobro u proseku, ali može da napravi katastrofu za neke upite
- Više iteracija može da izazove *klizanje upita*.
- Zašto?

Pseudo-RF i TREC4

- Cornell SMART sistem

Pseudo-RF i TREC4

- Cornell SMART sistem
- Rezultat prikazuje broj relevantnih dokumenata među najboljih 100 za 50 upita (ukupno 5000 dokumenata):

metod	broj relevantnih
lnc.ltc	3210
lnc.ltc-PsRF	3634
Lnu.ltu	3709
Lnu.ltu-PsRF	4350

Pseudo-RF i TREC4

- Cornell SMART sistem
- Rezultat prikazuje broj relevantnih dokumenata među najboljih 100 za 50 upita (ukupno 5000 dokumenata):

metod	broj relevantnih
lnc.ltc	3210
lnc.ltc-PsRF	3634
Lnu.ltu	3709
Lnu.ltu-PsRF	4350

- Porede se dva načina za normalizaciju (L i l) i pseudo-RF (PsRF)

Pseudo-RF i TREC4

- Cornell SMART sistem
- Rezultat prikazuje broj relevantnih dokumenata među najboljih 100 za 50 upita (ukupno 5000 dokumenata):

metod	broj relevantnih
lnc.ltc	3210
lnc.ltc-PsRF	3634
Lnu.ltu	3709
Lnu.ltu-PsRF	4350

- Porede se dva načina za normalizaciju (L i l) i pseudo-RF (PsRF)
- Pseudo-RF je dodao samo 20 termova upitu (Rocchio bi dodao mnogo više)

Pseudo-RF i TREC4

- Cornell SMART sistem
- Rezultat prikazuje broj relevantnih dokumenata među najboljih 100 za 50 upita (ukupno 5000 dokumenata):

metod	broj relevantnih
lnc.ltc	3210
lnc.ltc-PsRF	3634
Lnu.ltu	3709
Lnu.ltu-PsRF	4350

- Porede se dva načina za normalizaciju (L i l) i pseudo-RF (PsRF)
- Pseudo-RF je dodao samo 20 termova upitu (Rocchio bi dodao mnogo više)
- Ovo pokazuje da je pseudo-RF efikasan u proseku

Indirektan ili implicitni RF

- Od korisnika se ne očekuje eksplicitno izjašnjavanje da li je nešto relevantno ili nije, ali se koristi implicitno izjašnjavanje korisnika

Indirektan ili implicitni RF

- Od korisnika se ne očekuje eksplicitno izjašnjavanje da li je nešto relevantno ili nije, ali se koristi implicitno izjašnjavanje korisnika
- Na primer, ako se za jedan upit na veb pretraživaču na osnovu dinamičkog sažetka korisnici često odlučuju da otvore određeni rezultat, veb pretraživači mogu koristiti reči iz tog dinamičkog sažetka da njime prošire inicijalni upit i vrate listu odgovora za prošireni upit

Globalno proširenje upita

- Proširenje upita je drugi metod za povećanje povrata

Globalno proširenje upita

- Proširenje upita je drugi metod za povećanje povrata
- Koristimo termin „globalno proširenje upita“ kada mislimo na „globalne metode za reformulaciju upita“

Globalno proširenje upita

- Proširenje upita je drugi metod za povećanje povrata
- Koristimo termin „globalno proširenje upita“ kada mislimo na „globalne metode za reformulaciju upita“
- U GPU, upit se menja zavisno od nekog globalnog resursa, koji ne zavisi od upita

Globalno proširenje upita

- Proširenje upita je drugi metod za povećanje povrata
- Koristimo termin „globalno proširenje upita“ kada mislimo na „globalne metode za reformulaciju upita“
- U GPU, upit se menja zavisno od nekog globalnog resursa, koji ne zavisi od upita
- Glavne informacije koje se koriste: (skoro-)sinonimi

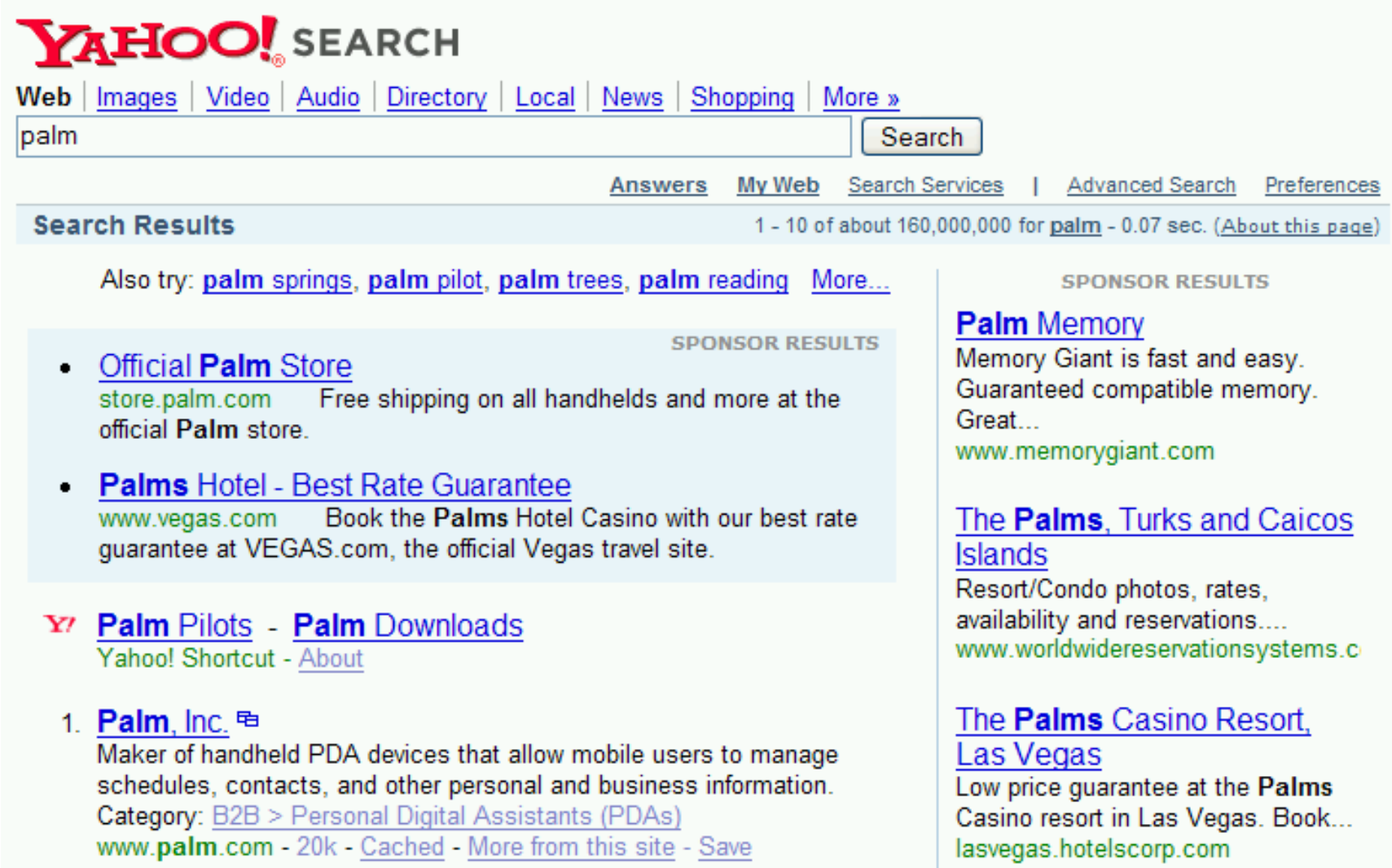
Globalno proširenje upita

- Proširenje upita je drugi metod za povećanje povrata
- Koristimo termin „globalno proširenje upita“ kada mislimo na „globalne metode za reformulaciju upita“
- U GPU, upit se menja zavisno od nekog globalnog resursa, koji ne zavisi od upita
- Glavne informacije koje se koriste: (skoro-)sinonimi
- Baza podataka koja čuva (skoro-)sinonime je tezaurus

Globalno proširenje upita

- Proširenje upita je drugi metod za povećanje povrata
- Koristimo termin „globalno proširenje upita“ kada mislimo na „globalne metode za reformulaciju upita“
- U GPU, upit se menja zavisno od nekog globalnog resursa, koji ne zavisi od upita
- Glavne informacije koje se koriste: (skoro-)sinonimi
- Baza podataka koja čuva (skoro-)sinonime je tezaurus
- Dve vrste tezaurusa: ručno i automatski formirani

Globalno proširenje upita: primer



YAHOO! SEARCH

Web | Images | Video | Audio | Directory | Local | News | Shopping | More »

palm Search

Answers My Web Search Services | Advanced Search Preferences

Search Results 1 - 10 of about 160,000,000 for palm - 0.07 sec. (About this page)

Also try: **palm** springs, **palm** pilot, **palm** trees, **palm** reading More...

SPONSOR RESULTS

- Official **Palm** Store
  store.palm.com Free shipping on all handhelds and more at the official **Palm** store.
- **Palms** Hotel - Best Rate Guarantee
  www.vegas.com Book the **Palms** Hotel Casino with our best rate guarantee at VEGAS.com, the official Vegas travel site.

Y! **Palm** Pilots - **Palm** Downloads
Yahoo! Shortcut - About

1. **Palm**, Inc.
   Maker of handheld PDA devices that allow mobile users to manage schedules, contacts, and other personal and business information.
   Category: B2B > Personal Digital Assistants (PDAs)
   www.**palm**.com - 20k - Cached - More from this site - Save

SPONSOR RESULTS

Palm Memory
Memory Giant is fast and easy. Guaranteed compatible memory. Great...
www.memorygiant.com

The **Palms**, Turks and Caicos Islands
Resort/Condo photos, rates, availability and reservations....
www.worldwidereservationsystems.c

The **Palms** Casino Resort, Las Vegas
Low price guarantee at the **Palms** Casino resort in Las Vegas. Book...
lasvegas.hotelscorp.com

Vrste odgovora korisnika

- Korisnik daje odgovor (informacije) o dokumentima.

Vrste odgovora korisnika

- Korisnik daje odgovor (informacije) o dokumentima.
  - Češće u RF

Vrste odgovora korisnika

- Korisnik daje odgovor (informacije) o dokumentima.
  - Češće u RF
- Korisnik daje odgovor o rečima ili frazama.

Vrste odgovora korisnika

- Korisnik daje odgovor (informacije) o dokumentima.
  - Češće u RF
- Korisnik daje odgovor o rečima ili frazama.
  - češće za proširenje upita

Vrste odgovora korisnika

- Korisnik daje odgovor (informacije) o dokumentima.
  - Češće u RF
- Korisnik daje odgovor o rečima ili frazama.
  - češće za proširenje upita
- RF možemo posmatrati kao varijantu proširenja upita

Vrste odgovora korisnika

- Korisnik daje odgovor (informacije) o dokumentima.
  - Češće u RF
- Korisnik daje odgovor o rečima ili frazama.
  - češće za proširenje upita
- RF možemo posmatrati kao varijantu proširenja upita
- Dodajemo termove u upit

Vrste odgovora korisnika

- Korisnik daje odgovor (informacije) o dokumentima.
  - Češće u RF
- Korisnik daje odgovor o rečima ili frazama.
  - češće za proširenje upita
- RF možemo posmatrati kao varijantu proširenja upita
- Dodajemo termove u upit
- Termovi dodati u RF ciklusu su zasnovani na „lokalnim“ informacijama u tekućem rezultatu

Vrste odgovora korisnika

- Korisnik daje odgovor (informacije) o dokumentima.
  - Češće u RF
- Korisnik daje odgovor o rečima ili frazama.
  - češće za proširenje upita
- RF možemo posmatrati kao varijantu proširenja upita
- Dodajemo termove u upit
- Termovi dodati u RF ciklusu su zasnovani na „lokalnim“ informacijama u tekućem rezultatu
- Termovi dodati u GPU se zasnivaju na globalnim informacijama koje ne zavise od upita

Vrste proširenja upita

- Ručni tezaurus (neko ga održava, npr. PubMed)

Vrste proširenja upita

- Ručni tezaurus (neko ga održava, npr. PubMed)
- Automatski generisan tezaurus (npr. zasnovan na statistici zajedničkog pojavljivanja)

Vrste proširenja upita

- Ručni tezaurus (neko ga održava, npr. PubMed)
- Automatski generisan tezaurus (npr. zasnovan na statistici zajedničkog pojavljivanja)
- Ekvivalentnost upita zasnovano na analizi istorije upita (log mining, korisno za web)

Proširenje upita zasnovano na tezaurusu

- Za svaki term t u upitu, proširi upit rečima koje tezaurus navodi kao semantički povezane sa t

Proširenje upita zasnovano na tezaurusu

- Za svaki term t u upitu, proširi upit rečima koje tezaurus navodi kao semantički povezane sa t
- Primer: $\text{hospital} \rightarrow \text{medical}$

Proširenje upita zasnovano na tezaurusu

- Za svaki term t u upitu, proširi upit rečima koje tezaurus navodi kao semantički povezane sa t
- Primer: $\text{hospital} \rightarrow \text{medical}$
- U principu povećava povrat

Proširenje upita zasnovano na tezaurusu

- Za svaki term t u upitu, proširi upit rečima koje tezaurus navodi kao semantički povezane sa t
- Primer: $\text{hospital} \rightarrow \text{medical}$
- U principu povećava povrat
- Može značajno da smanji preciznost, naročito sa dvosmislenim termovima

Proširenje upita zasnovano na tezaurusu

- Za svaki term t u upitu, proširi upit rečima koje tezaurus navodi kao semantički povezane sa t
- Primer: hospital $\rightarrow$ medical
- U principu povećava povrat
- Može značajno da smanji preciznost, naročito sa dvosmislenim termovima
  - interest rate (kamatna stopa) $\rightarrow$ interest rate fascinate evaluate

Proširenje upita zasnovano na tezaurusu

- Za svaki term t u upitu, proširi upit rečima koje tezaurus navodi kao semantički povezane sa t
- Primer: $\text{hospital} \rightarrow \text{medical}$
- U principu povećava povrat
- Može značajno da smanji preciznost, naročito sa dvosmislenim termovima
  - interest rate (kamatna stopa) $\rightarrow$ $\text{interest rate fascinate evaluate}$
- Često korišćen u specijalizovanim pretraživačima za naučne i inženjerske primene

Proširenje upita zasnovano na tezaurusu

- Za svaki term t u upitu, proširi upit rečima koje tezaurus navodi kao semantički povezane sa t
- Primer: $\text{hospital} \rightarrow \text{medical}$
- U principu povećava povrat
- Može značajno da smanji preciznost, naročito sa dvosmislenim termovima
  - interest rate (kamatna stopa) $\rightarrow \text{interest rate fascinate evaluate}$
- Često korišćen u specijalizovanim pretraživačima za naučne i inženjerske primene
- Vrlo skupo ručno održavanje tezaurusa

Proširenje upita zasnovano na tezaurusu

- Za svaki term t u upitu, proširi upit rečima koje tezaurus navodi kao semantički povezane sa t
- Primer: $\text{hospital} \rightarrow \text{medical}$
- U principu povećava povrat
- Može značajno da smanji preciznost, naročito sa dvosmislenim termovima
  - interest rate (kamatna stopa) $\rightarrow$ $\text{interest rate fascinate evaluate}$
- Često korišćen u specijalizovanim pretraživačima za naučne i inženjerske primene
- Vrlo skupo ručno održavanje tezaurusa
- Ručni tezaurus je skoro ekvivalentan anotiranju sa *kontrolisanim rečnikom*.

Primer ručnog tezaurusa: PubMed

Automatsko generisanje tezaurusa

- Pokušaj da se generiše tezaurus na osnovu analize distribucije reči u dokumentu

Automatsko generisanje tezaurusa

- Pokušaj da se generiše tezaurus na osnovu analize distribucije reči u dokumentu
- Osnovni pojam: sličnost dve reči

Automatsko generisanje tezaurusa

- Pokušaj da se generiše tezaurus na osnovu analize distribucije reči u dokumentu
- Osnovni pojam: sličnost dve reči
- Definicija 1: dve reči su slične ako se „pojavljuju zajedno“ sa sličnim rečima

Automatsko generisanje tezaurusa

- Pokušaj da se generiše tezaurus na osnovu analize distribucije reči u dokumentu
- Osnovni pojam: sličnost dve reči
- Definicija 1: dve reči su slične ako se „pojavljuju zajedno“ sa sličnim rečima
- Definicija 2: dve reči su slične ako se pojavljuju u istom gramatičkom odnosu sa sličnim rečima

Automatsko generisanje tezaurusa

- Pokušaj da se generiše tezaurus na osnovu analize distribucije reči u dokumentu
- Osnovni pojam: sličnost dve reči
- Definicija 1: dve reči su slične ako se „pojavljuju zajedno“ sa sličnim rečima
- Definicija 2: dve reči su slične ako se pojavljuju u istom gramatičkom odnosu sa sličnim rečima
  - Jabuke i kruške se beru, ljušte, jedu, pripremaju; dakle jabuke i kruške su slične

Automatsko generisanje tezaurusa

- Pokušaj da se generiše tezaurus na osnovu analize distribucije reči u dokumentu
- Osnovni pojam: sličnost dve reči
- Definicija 1: dve reči su slične ako se „pojavljuju zajedno“ sa sličnim rečima
- Definicija 2: dve reči su slične ako se pojavljuju u istom gramatičkom odnosu sa sličnim rečima
  - Jabuke i kruške se beru, ljušte, jedu, pripremaju; dakle jabuke i kruške su slične
- Zajedničko pojavljivanje je robusnije, gramatički odnosi su tačniji

Tezarus baziran na zajedničkom pojavljivanju: primeri

reč	najbliži susedi
absolutely	absurd, whatsoever, totally, exactly, nothing
bottomed	dip, copper, drops, topped, slide, trimmed
captivating	shimmer, stunningly, superbly, plucky, witty
doghouse	dog, porch, crawling, beside, downstairs
makeup	repellent, lotion, glossy, sunscreen, skin, gel
mediating	reconciliation, negotiate, case, conciliation
keeping	hoping, bring, wiping, could, some, would
lithographs	drawings, Picasso, Dali, sculptures, Gauguin
pathogens	toxins, bacteria, organisms, bacterial, parasite
senses	grasp, psyche, truly, clumsy, naive, innate

Rezime

- RF i proširenje upita povećavaju povrat

Rezime

- RF i proširenje upita povećavaju povrat
- U puno slučajeva preciznost značajno opada

Rezime

- RF i proširenje upita povećavaju povrat
- U puno slučajeva preciznost značajno opada
- Izmena upita zasnovana na istoriji upita (query log) (koja je složenija od jednostavnog proširenja) je češća na webu od RF-a