Performanse IR sistema

Dragan Ivanović
dragan.ivanovic@uns.ac.rs

Katedra za informatiku, Fakultet tehničkih nauka, Novi Sad

2015.

Mere za kvalitet pretraživača

- Koliko brzo indeksira

Mere za kvalitet pretraživača

- Koliko brzo indeksira
  - Broj dokumenata/megabajta na sat

Mere za kvalitet pretraživača

- Koliko brzo indeksira
  - Broj dokumenata/megabajta na sat
- Koliko brzo pretražuje

Mere za kvalitet pretraživača

- Koliko brzo indeksira
  - Broj dokumenata/megabajta na sat
- Koliko brzo pretražuje
  - Kašnjenje kao funkcija veličine indeksa i broja upita u sekundi

Mere za kvalitet pretraživača

- Svi prethodni kriterijumi su merljivi: možemo kvantifikovati brzinu / prostor / novac

Mere za kvalitet pretraživača

- Svi prethodni kriterijumi su merljivi: možemo kvantifikovati brzinu / prostor / novac
- Međutim, ključna mera za pretraživač je zadovoljstvo korisnika

Ko je korisnik?

- Ko je korisnik koga želimo da zadovoljimo?

Ko je korisnik?

- Ko je korisnik koga želimo da zadovoljimo?
- Web pretraživači: *tragač*. Tragač pronalazi ono što traži.

Ko je korisnik?

- Ko je korisnik koga želimo da zadovoljimo?
- Web pretraživači: *tragač*. Tragač pronalazi ono što traži.

- Web pretraživači: zakupac reklama. Da li tragači klikću na moje reklame?

Ko je korisnik?

- Ko je korisnik koga želimo da zadovoljimo?
- Web pretraživači: *tragač*. Tragač pronalazi ono što traži.

- Web pretraživači: zakupac reklama. Da li tragači klikću na moje reklame?
- E-poslovanje: kupac. Kupac kupuje ono zbog čega je došao na sajt.

Ko je korisnik?

- Ko je korisnik koga želimo da zadovoljimo?
- Web pretraživači: *tragač*. Tragač pronalazi ono što traži.

- Web pretraživači: zakupac reklama. Da li tragači klikću na moje reklame?
- E-poslovanje: kupac. Kupac kupuje ono zbog čega je došao na sajt.

- E-poslovanje: prodavac. Prodavac može da prodaje svoju robu (jer je pretraživač uputio kupce na prave sadržaje).

Ko je korisnik?

- Ko je korisnik koga želimo da zadovoljimo?
- Web pretraživači: *tragač*. Tragač pronalazi ono što traži.

- Web pretraživači: zakupac reklama. Da li tragači klikću na moje reklame?
- E-poslovanje: kupac. Kupac kupuje ono zbog čega je došao na sajt.

- E-poslovanje: prodavac. Prodavac može da prodaje svoju robu (jer je pretraživač uputio kupce na prave sadržaje).

- Firma: direktor. Zaposleni su produktivniji jer brzo pronalaze ono što im treba.

Ko je korisnik?

- Ko je korisnik koga želimo da zadovoljimo?
- Web pretraživači: *tragač*. Tragač pronalazi ono što traži. Mera: stepen vraćanja na ovaj pretraživač
- Web pretraživači: zakupac reklama. Da li tragači klikću na moje reklame?
- E-poslovanje: kupac. Kupac kupuje ono zbog čega je došao na sajt.

- E-poslovanje: prodavac. Prodavac može da prodaje svoju robu (jer je pretraživač uputio kupce na prave sadržaje).

- Firma: direktor. Zaposleni su produktivniji jer brzo pronalaze ono što im treba.

Ko je korisnik?

- Ko je korisnik koga želimo da zadovoljimo?
- Web pretraživači: *tragač*. Tragač pronalazi ono što traži. Mera: stepen vraćanja na ovaj pretraživač
- Web pretraživači: zakupac reklama. Da li tragači klikću na moje reklame? Mera: clickthrough rate
- E-poslovanje: kupac. Kupac kupuje ono zbog čega je došao na sajt.

- E-poslovanje: prodavac. Prodavac može da prodaje svoju robu (jer je pretraživač uputio kupce na prave sadržaje).

- Firma: direktor. Zaposleni su produktivniji jer brzo pronalaze ono što im treba.

Ko je korisnik?

- Ko je korisnik koga želimo da zadovoljimo?
- Web pretraživači: *tragač*. Tragač pronalazi ono što traži. Mera: stepen vraćanja na ovaj pretraživač
- Web pretraživači: zakupac reklama. Da li tragači klikću na moje reklame? Mera: clickthrough rate
- E-poslovanje: kupac. Kupac kupuje ono zbog čega je došao na sajt. Mere: vreme do kupovine, procenat konvertovanih tragača u kupce
- E-poslovanje: prodavac. Prodavac može da prodaje svoju robu (jer je pretraživač uputio kupce na prave sadržaje).

- Firma: direktor. Zaposleni su produktivniji jer brzo pronalaze ono što im treba.

Ko je korisnik?

- Ko je korisnik koga želimo da zadovoljimo?
- Web pretraživači: *tragač*. Tragač pronalazi ono što traži. Mera: stepen vraćanja na ovaj pretraživač
- Web pretraživači: zakupac reklama. Da li tragači klikću na moje reklame? Mera: clickthrough rate
- E-poslovanje: kupac. Kupac kupuje ono zbog čega je došao na sajt. Mere: vreme do kupovine, procenat konvertovanih tragača u kupce
- E-poslovanje: prodavac. Prodavac može da prodaje svoju robu (jer je pretraživač uputio kupce na prave sadržaje). Mera: profit po prodatom artiklu
- Firma: direktor. Zaposleni su produktivniji jer brzo pronalaze ono što im treba.

Ko je korisnik?

- Ko je korisnik koga želimo da zadovoljimo?
- Web pretraživači: *tragač*. Tragač pronalazi ono što traži. Mera: stepen vraćanja na ovaj pretraživač
- Web pretraživači: zakupac reklama. Da li tragači klikću na moje reklame? Mera: clickthrough rate
- E-poslovanje: kupac. Kupac kupuje ono zbog čega je došao na sajt. Mere: vreme do kupovine, procenat konvertovanih tragača u kupce
- E-poslovanje: prodavac. Prodavac može da prodaje svoju robu (jer je pretraživač uputio kupce na prave sadržaje). Mera: profit po prodatom artiklu
- Firma: direktor. Zaposleni su produktivniji jer brzo pronalaze ono što im treba. Mera: profit firme

Zadovoljstvo korisnika

- Šta je zadovoljstvo korisnika?

Zadovoljstvo korisnika

- Šta je zadovoljstvo korisnika?
- Faktori zadovoljstva uključuju:

Zadovoljstvo korisnika

- Šta je zadovoljstvo korisnika?
- Faktori zadovoljstva uključuju:
  - Brzinu dobijanja odgovora

Zadovoljstvo korisnika

- Šta je zadovoljstvo korisnika?
- Faktori zadovoljstva uključuju:
  - Brzinu dobijanja odgovora
  - Veličinu indeksa

Zadovoljstvo korisnika

- Šta je zadovoljstvo korisnika?
- Faktori zadovoljstva uključuju:
  - Brzinu dobijanja odgovora
  - Veličinu indeksa
  - Nezatrpan korisnički interfejs

Zadovoljstvo korisnika

- Šta je zadovoljstvo korisnika?
- Faktori zadovoljstva uključuju:
  - Brzinu dobijanja odgovora
  - Veličinu indeksa
  - Nezatrpan korisnički interfejs
  - Najvažnije: relevantnost

Zadovoljstvo korisnika

- Šta je zadovoljstvo korisnika?
- Faktori zadovoljstva uključuju:
  - Brzinu dobijanja odgovora
  - Veličinu indeksa
  - Nezatrpan korisnički interfejs
  - Najvažnije: relevantnost
  - (Možda najvažnije: besplatan pristup)

Zadovoljstvo korisnika

- Šta je zadovoljstvo korisnika?
- Faktori zadovoljstva uključuju:
  - Brzinu dobijanja odgovora
  - Veličinu indeksa
  - Nezatrpan korisnički interfejs
  - Najvažnije: relevantnost
  - (Možda najvažnije: besplatan pristup)
- Nijedan faktor pojedinačno nije dovoljan: fantastično brzi ali beskorisni odgovori neće korisnika učiniti zadovoljnim

Zadovoljstvo korisnika

- Šta je zadovoljstvo korisnika?
- Faktori zadovoljstva uključuju:
  - Brzinu dobijanja odgovora
  - Veličinu indeksa
  - Nezatrpan korisnički interfejs
  - Najvažnije: relevantnost
  - (Možda najvažnije: besplatan pristup)
- Nijedan faktor pojedinačno nije dovoljan: fantastično brzi ali beskorisni odgovori neće korisnika učiniti zadovoljnim
- Kako da kvanitifikujemo zadovoljstvo korisnika?

Najčešća definicija zadovoljstva korisnika: relevantnost

- Zadovoljstvo korisnika se izjednačava sa relevantnošću rezultata pretrage

Najčešća definicija zadovoljstva korisnika: relevantnost

- Zadovoljstvo korisnika se izjednačava sa relevantnošću rezultata pretrage
- Kako meriti relevantnost?

Najčešća definicija zadovoljstva korisnika: relevantnost

- Zadovoljstvo korisnika se izjednačava sa relevantnošću rezultata pretrage
- Kako meriti relevantnost?
- Standardna metodologija u IR ima tri elementa

Najčešća definicija zadovoljstva korisnika: relevantnost

- Zadovoljstvo korisnika se izjednačava sa relevantnošću rezultata pretrage
- Kako meriti relevantnost?
- Standardna metodologija u IR ima tri elementa
  - test-kolekciju dokumenata

Najčešća definicija zadovoljstva korisnika: relevantnost

- Zadovoljstvo korisnika se izjednačava sa relevantnošću rezultata pretrage
- Kako meriti relevantnost?
- Standardna metodologija u IR ima tri elementa
  - test-kolekciju dokumenata
  - skup test-upita

Najčešća definicija zadovoljstva korisnika: relevantnost

- Zadovoljstvo korisnika se izjednačava sa relevantnošću rezultata pretrage
- Kako meriti relevantnost?
- Standardna metodologija u IR ima tri elementa
  - test-kolekciju dokumenata
  - skup test-upita
  - binarnu (ili, ređe, ne-binarnu) ocenu relevatnosti svakog para upit-dokument

Najčešća definicija zadovoljstva korisnika: relevantnost

- Zadovoljstvo korisnika se izjednačava sa relevantnošću rezultata pretrage
- Kako meriti relevantnost?
- Standardna metodologija u IR ima tri elementa
  - test-kolekciju dokumenata
  - skup test-upita
  - binarnu (ili, ređe, ne-binarnu) ocenu relevatnosti svakog para upit-dokument
- Ovakvo vrednovanje (veštački scenariji) se često kritikuje

Najčešća definicija zadovoljstva korisnika: relevantnost

- Zadovoljstvo korisnika se izjednačava sa relevantnošću rezultata pretrage
- Kako meriti relevantnost?
- Standardna metodologija u IR ima tri elementa
  - test-kolekciju dokumenata
  - skup test-upita
  - binarnu (ili, ređe, ne-binarnu) ocenu relevatnosti svakog para upit-dokument
- Ovakvo vrednovanje (veštački scenariji) se često kritikuje
- Ali je vrlo uspešno u IR

Relevantnost: upit ili potreba za informacijama

- Relevantnost u odnosu na šta?

Relevantnost: upit ili potreba za informacijama

- Relevantnost u odnosu na šta?
- Proba 1: relevantnost u odnosu na upit

Relevantnost: upit ili potreba za informacijama

- Relevantnost u odnosu na šta?
- Proba 1: relevantnost u odnosu na upit
- „Relevantnost u odnosu na upit“ je vrlo problematična

Relevantnost: upit ili potreba za informacijama

- Relevantnost u odnosu na šta?
- Proba 1: relevantnost u odnosu na upit
- „Relevantnost u odnosu na upit“ je vrlo problematična
- Informaciona potreba *i*: Tražimo informacije o tome da li je crno vino bolje za smanjenje rizika od infarkta nego belo vino

Relevantnost: upit ili potreba za informacijama

- Relevantnost u odnosu na šta?
- Proba 1: relevantnost u odnosu na upit
- „Relevantnost u odnosu na upit“ je vrlo problematična
- Informaciona potreba *i*: Tražimo informacije o tome da li je crno vino bolje za smanjenje rizika od infarkta nego belo vino
- Ovo je informaciona potreba, a ne upit

Relevantnost: upit ili potreba za informacijama

- Relevantnost u odnosu na šta?
- Proba 1: relevantnost u odnosu na upit
- „Relevantnost u odnosu na upit“ je vrlo problematična
- Informaciona potreba i: Tražimo informacije o tome da li je crno vino bolje za smanjenje rizika od infarkta nego belo vino
- Ovo je informaciona potreba, a ne upit
- Upit q: wine AND red AND white AND heart AND attack

Relevantnost: upit ili potreba za informacijama

- Relevantnost u odnosu na šta?
- Proba 1: relevantnost u odnosu na upit
- „Relevantnost u odnosu na upit“ je vrlo problematična
- Informaciona potreba i: Tražimo informacije o tome da li je crno vino bolje za smanjenje rizika od infarkta nego belo vino
- Ovo je informaciona potreba, a ne upit
- Upit q: wine AND red AND white AND heart AND attack
- Razmotrimo dokument d': *He then launched into the heart of his speech and attacked the wine industry lobby for downplaying the role of red and white wine in drunk driving.*

Relevantnost: upit ili potreba za informacijama

- Relevantnost u odnosu na šta?
- Proba 1: relevantnost u odnosu na upit
- „Relevantnost u odnosu na upit“ je vrlo problematična
- Informaciona potreba i: Tražimo informacije o tome da li je crno vino bolje za smanjenje rizika od infarkta nego belo vino
- Ovo je informaciona potreba, a ne upit
- Upit q: wine AND red AND white AND heart AND attack
- Razmotrimo dokument d': *He then launched into the heart of his speech and attacked the wine industry lobby for downplaying the role of red and white wine in drunk driving.*
- d' je relevantan za upit q ...

Relevantnost: upit ili potreba za informacijama

- Relevantnost u odnosu na šta?
- Proba 1: relevantnost u odnosu na upit
- „Relevantnost u odnosu na upit“ je vrlo problematična
- Informaciona potreba i: Tražimo informacije o tome da li je crno vino bolje za smanjenje rizika od infarkta nego belo vino
- Ovo je informaciona potreba, a ne upit
- Upit q: wine AND red AND white AND heart AND attack
- Razmotrimo dokument d': *He then launched into the heart of his speech and attacked the wine industry lobby for downplaying the role of red and white wine in drunk driving.*
- d' je relevantan za upit q ...
- d' nije relevantan za informacionu potrebu i.

Relevantnost: upit ili potreba za informacijama

- Zadovoljstvo korisnika se može meriti samo prema relevantnosti u odnosu na informacione potrebe, a ne upite
- Terminologija nam je aljkava: govorimo o relevantnosti upit-dokument kada mislimo na relevantnost informaciona potreba-dokument

Kritika čiste relevantnosti

- Definisana je relevantnost za izolovani par upit-dokument

Kritika čiste relevantnosti

- Definisana je relevantnost za izolovani par upit-dokument
- Alternativna definicija: marginalna relevantnost

Kritika čiste relevantnosti

- Definisana je relevantnost za izolovani par upit-dokument
- Alternativna definicija: marginalna relevantnost
- Marginalna relevantnost dokumenta u rezultatu je dodatna informacija koju dokument donosi

Kritika čiste relevantnosti

- Definisana je relevantnost za izolovani par upit-dokument
- Alternativna definicija: marginalna relevantnost
- Marginalna relevantnost dokumenta u rezultatu je dodatna informacija koju dokument donosi
- Primer: duplikat može biti vrlo relevantan ali ima marginalnu relevantnost 0

Kritika čiste relevantnosti

- Definisana je relevantnost za izolovani par upit-dokument
- Alternativna definicija: marginalna relevantnost
- Marginalna relevantnost dokumenta u rezultatu je dodatna informacija koju dokument donosi
- Primer: duplikat može biti vrlo relevantan ali ima marginalnu relevantnost 0
- Marginalna relevantnost je bolja mera zadovoljstva korisnika

Kritika čiste relevantnosti

- Definisana je relevantnost za izolovani par upit-dokument
- Alternativna definicija: marginalna relevantnost
- Marginalna relevantnost dokumenta u rezultatu je dodatna informacija koju dokument donosi
- Primer: duplikat može biti vrlo relevantan ali ima marginalnu relevantnost 0
- Marginalna relevantnost je bolja mera zadovoljstva korisnika
- Ali je praktično nemoguće sprovoditi eksperimente bazirane na marginalnoj relevantnosti

Precision/recall – preciznost/povrat

- Preciznost P je deo pronađenih dokumenata koji su relevantni

$$\text{Preciznost} = \frac{\#(\text{pronađeni relevantni})}{\#(\text{svi pronađeni})} = P(\text{relevantan}|\text{pronađen})$$

Precision/recall – preciznost/povrat

- Preciznost P je deo pronađenih dokumenata koji su relevantni

$$\text{Preciznost} = \frac{\#(\text{pronađeni relevantni})}{\#(\text{svi pronađeni})} = P(\text{relevantan}|\text{pronađen})$$

- Povrat R je deo relevantnih dokumenata koji su pronađeni

$$\text{Povrat} = \frac{\#(\text{pronađeni relevatni})}{\#(\text{svi relevantni})} = P(\text{pronađen}|\text{relevantan})$$

Precision/recall – preciznost/povrat

- Preciznost P je deo pronađenih dokumenata koji su relevantni

$$\text{Preciznost} = \frac{\#(\text{pronađeni relevantni})}{\#(\text{svi pronađeni})} = P(\text{relevantan}|\text{pronađen})$$

- Povrat R je deo relevantnih dokumenata koji su pronađeni

$$\text{Povrat} = \frac{\#(\text{pronađeni relevatni})}{\#(\text{svi relevantni})} = P(\text{pronađen}|\text{relevantan})$$

Precision/recall – preciznost/povrat

	Relevantan	Nerelevantan
Pronađen	true positives (TP)	false positives (FP)
Nije pronađen	false negatives (FN)	true negatives (TN)

$$P = \frac{TP}{TP + FP}$$

$$R = \frac{TP}{TP + FN}$$

Tačnost

- Zašto koristimo složene mere kao preciznost i povrat?

Tačnost

- Zašto koristimo složene mere kao preciznost i povrat?
- Zašto ne nešto jednostavno, npr. tačnost?

Tačnost

- Zašto koristimo složene mere kao preciznost i povrat?
- Zašto ne nešto jednostavno, npr. tačnost?
- Tačnost je deo odluka (relevantan/nerelevantan) koje su ispravne

Tačnost

- Zašto koristimo složene mere kao preciznost i povrat?
- Zašto ne nešto jednostavno, npr. tačnost?
- Tačnost je deo odluka (relevantan/nerelevantan) koje su ispravne
- U smislu prethodne tabele, tačnost $= \frac{TP+TN}{TP+FP+FN+TN}$.

Tačnost

- Zašto koristimo složene mere kao preciznost i povrat?
- Zašto ne nešto jednostavno, npr. tačnost?
- Tačnost je deo odluka (relevantan/nerelevantan) koje su ispravne
- U smislu prethodne tabele, tačnost $= \frac{TP+TN}{TP+FP+FN+TN}$.
- Zašto tačnost nije korisna mera za web IR?

- Jednostavan štos za maksimizaciju tačnosti u IR: uvek kaži ne i vrati prazan skup

- Jednostavan štos za maksimizaciju tačnosti u IR: uvek kaži ne i vrati prazan skup
- Imaćeš 99.99% tačnost za većinu upita

- Jednostavan štos za maksimizaciju tačnosti u IR: uvek kaži ne i vrati prazan skup
- Imaćeš 99.99% tačnost za većinu upita
- Tragači na webu (i u IR uopšte) žele da pronađu nešto i imaju određeni stepen tolerancije na đubre

- Jednostavan štos za maksimizaciju tačnosti u IR: uvek kaži ne i vrati prazan skup
- Imaćeš 99.99% tačnost za većinu upita
- Tragači na webu (i u IR uopšte) žele da pronađu nešto i imaju određeni stepen tolerancije na đubre
- Tačnost nije dobra mera zadovoljstva korisnika, pa ćemo koristiti preciznost i povrat

Teškoće u korišćenju precision/recall

- Treba nam ocena relevantnosti za parove informaciona potreba-dokument ali je njih teško/skupo napraviti

Teškoće u korišćenju precision/recall

- Treba nam ocena relevantnosti za parove informaciona potreba-dokument ali je njih teško/skupo napraviti
- Alternative korišćenju precision/recall i pravljenju ocena... na kraju predavanja

Precision/recall kompromis

- Može se povećati povrat vraćanjem više dokumenata

Precision/recall kompromis

- Može se povećati povrat vraćanjem više dokumenata
- Povrat je neopadajuća funkcija broja pronađenih dokumenata

Precision/recall kompromis

- Može se povećati povrat vraćanjem više dokumenata
- Povrat je neopadajuća funkcija broja pronađenih dokumenata
- Sistem koji vraća sve dokumente ima 100% povrat!

Precision/recall kompromis

- Može se povećati povrat vraćanjem više dokumenata
- Povrat je neopadajuća funkcija broja pronađenih dokumenata
- Sistem koji vraća sve dokumente ima 100% povrat!
- Suprotno je takođe tačno (često): lako je imati veliku preciznost za mali povrat

Precision/recall kompromis

- Može se povećati povrat vraćanjem više dokumenata
- Povrat je neopadajuća funkcija broja pronađenih dokumenata
- Sistem koji vraća sve dokumente ima 100% povrat!
- Suprotno je takođe tačno (često): lako je imati veliku preciznost za mali povrat
- Neka je najbolje rangirani dokument relevantan. Kako možemo maksimizovati preciznost?

Kombinovana mera: F

- F omogućava da merimo kompromis između preciznosti i povrata

Kombinovana mera: F

- F omogućava da merimo kompromis između preciznosti i povrata
-

$$F = \frac{1}{\alpha\frac{1}{P} + (1-\alpha)\frac{1}{R}} = \frac{(\beta^2+1)PR}{\beta^2 P + R} \quad \text{gde} \quad \beta^2 = \frac{1-\alpha}{\alpha}$$

Kombinovana mera: F

- F omogućava da merimo kompromis između preciznosti i povrata
- $$F = \frac{1}{\alpha\frac{1}{P} + (1-\alpha)\frac{1}{R}} = \frac{(\beta^2+1)PR}{\beta^2 P + R} \quad \text{gde} \quad \beta^2 = \frac{1-\alpha}{\alpha}$$
- $\alpha \in [0,1]$ pa prema tome $\beta^2 \in [0,\infty]$

Kombinovana mera: F

- F omogućava da merimo kompromis između preciznosti i povrata
-

$$F = \frac{1}{\alpha\frac{1}{P} + (1-\alpha)\frac{1}{R}} = \frac{(\beta^2+1)PR}{\beta^2 P + R} \quad \text{gde} \quad \beta^2 = \frac{1-\alpha}{\alpha}$$

- $\alpha \in [0, 1]$ pa prema tome $\beta^2 \in [0, \infty]$
- Najčešće korišćen: balansirani F sa $\beta = 1$ ili $\alpha = 0.5$

Kombinovana mera: F

- F omogućava da merimo kompromis između preciznosti i povrata
-

$$F = \frac{1}{\alpha\frac{1}{P} + (1-\alpha)\frac{1}{R}} = \frac{(\beta^2+1)PR}{\beta^2 P + R} \quad \text{gde} \quad \beta^2 = \frac{1-\alpha}{\alpha}$$

- $\alpha \in [0, 1]$ pa prema tome $\beta^2 \in [0, \infty]$
- Najčešće korišćen: balansirani F sa $\beta = 1$ ili $\alpha = 0.5$
  - Ovo je harmonijska sredina P i R: $\frac{1}{F} = \frac{1}{2}(\frac{1}{P} + \frac{1}{R})$

Kombinovana mera: F

- F omogućava da merimo kompromis između preciznosti i povrata
- $$F = \frac{1}{\alpha\frac{1}{P} + (1-\alpha)\frac{1}{R}} = \frac{(\beta^2+1)PR}{\beta^2 P + R} \quad \text{gde} \quad \beta^2 = \frac{1-\alpha}{\alpha}$$
- $\alpha \in [0,1]$ pa prema tome $\beta^2 \in [0,\infty]$
- Najčešće korišćen: balansirani F sa $\beta = 1$ ili $\alpha = 0.5$
  - Ovo je harmonijska sredina P i R: $\frac{1}{F} = \frac{1}{2}(\frac{1}{P} + \frac{1}{R})$
- Koji opseg vrednosti za β da izaberemo da povrat vrednujemo više nego preciznost?

Kombinovana mera: F

- F omogućava da merimo kompromis između preciznosti i povrata
-

$$F = \frac{1}{\alpha\frac{1}{P} + (1-\alpha)\frac{1}{R}} = \frac{(\beta^2+1)PR}{\beta^2 P + R} \quad \text{gde} \quad \beta^2 = \frac{1-\alpha}{\alpha}$$

- $\alpha \in [0,1]$ pa prema tome $\beta^2 \in [0,\infty]$
- Najčešće korišćen: balansirani F sa $\beta = 1$ ili $\alpha = 0.5$
  - Ovo je harmonijska sredina P i R: $\frac{1}{F} = \frac{1}{2}(\frac{1}{P} + \frac{1}{R})$
- Koji opseg vrednosti za β da izaberemo da povrat vrednujemo više nego preciznost?
- Kada je $\beta > 1$ povrat vrednujemo više nego preciznost, a kada je $\beta < 1$ onda preciznost vrednujemo više nego povrat

F: primer

	relevantni	nerelevantni
pronađeni	18	2
nepronađeni	82	1,000,000,000

F: primer

	relevantni	nerelevantni
pronađeni	18	2
nepronađeni	82	1,000,000,000

- preciznost?

F: primer

	relevantni	nerelevantni
pronađeni	18	2
nepronađeni	82	1,000,000,000

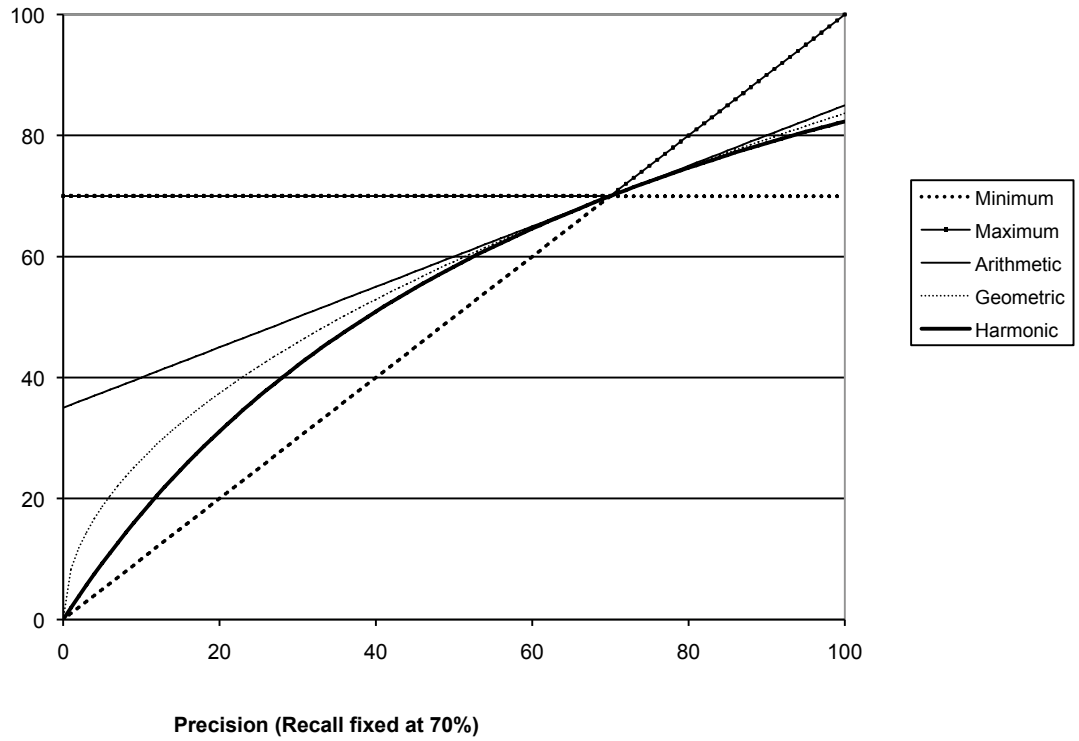
- povrat?

F: primer

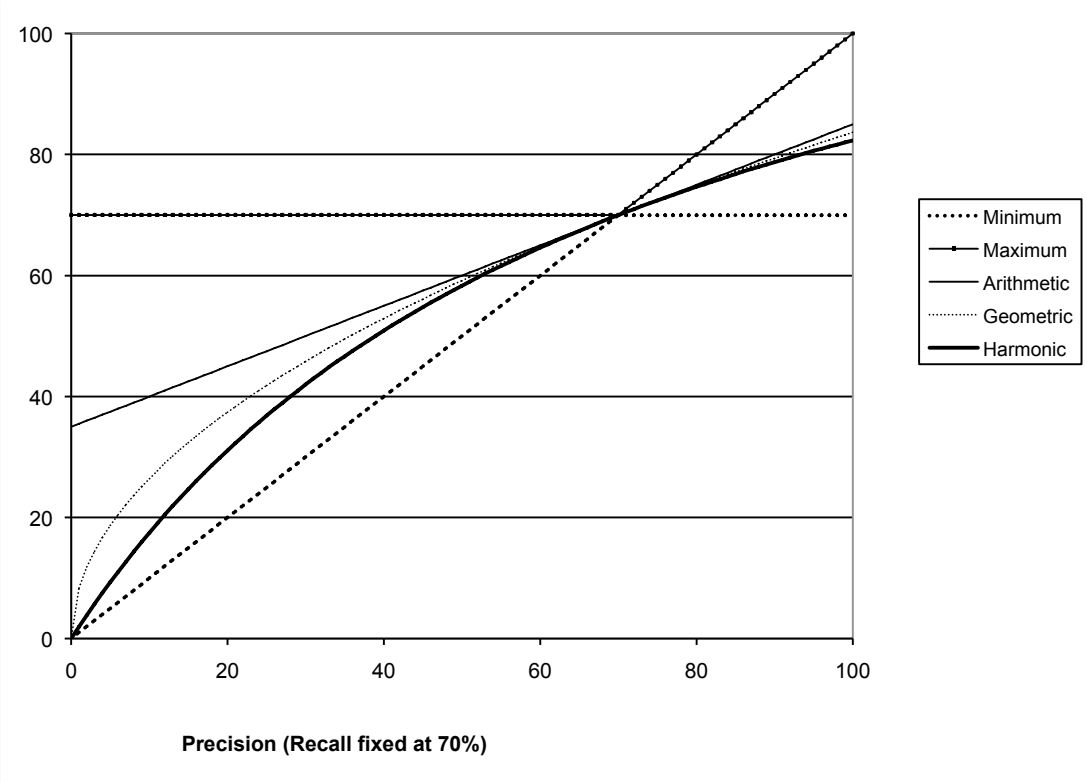
	relevantni	nerelevantni
pronađeni	18	2
nepronađeni	82	1,000,000,000

- F_1?

F_1 i druge mere



F_1 i druge mere



- Možemo posmatrati harmonijsku sredinu kao meki minimum

F: Zašto harmonijska sredina?

- Aritmetička sredina je 50% za pretraživač koji „vraća sve“ što je previše

F: Zašto harmonijska sredina?

- Aritmetička sredina je 50% za pretraživač koji „vraća sve“ što je previše
- Želja: kaznimo loše performanse na račun bilo preciznosti ili povrata

F: Zašto harmonijska sredina?

- Aritmetička sredina je 50% za pretraživač koji „vraća sve“ što je previše
- Želja: kaznimo loše performanse na račun bilo preciznosti ili povrata
- Ovo se postiže uzimanjem minimuma

F: Zašto harmonijska sredina?

- Aritmetička sredina je 50% za pretraživač koji „vraća sve“ što je previše
- Želja: kaznimo loše performanse na račun bilo preciznosti ili povrata
- Ovo se postiže uzimanjem minimuma
- F (harmonijska sredina) je kao meki minimum

Precision/recall kriva

- Preciznost/povrat/F su mere nerangiranih skupova.

Precision/recall kriva

- Preciznost/povrat/F su mere nerangiranih skupova.
- Lako ih možemo pretvoriti u mere rangiranih lista.

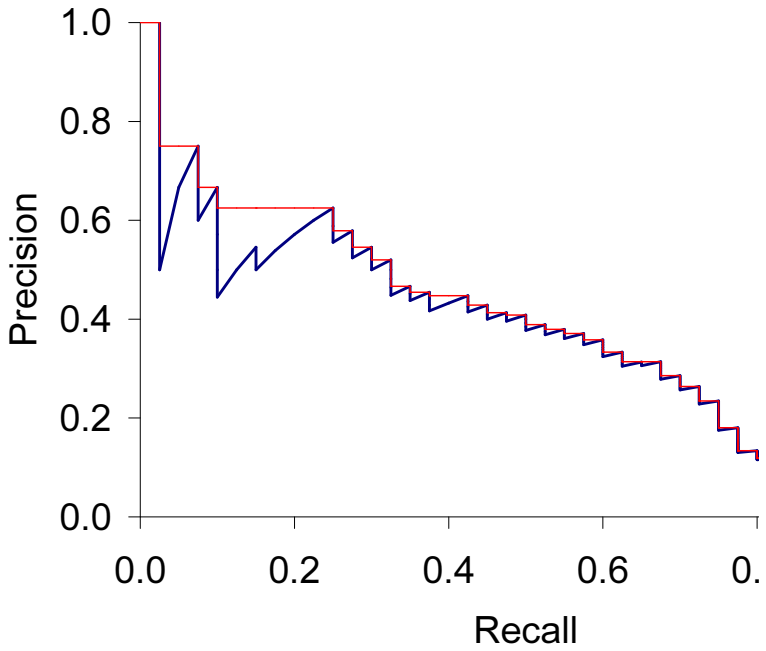
Precision/recall kriva

- Preciznost/povrat/F su mere nerangiranih skupova.
- Lako ih možemo pretvoriti u mere rangiranih lista.
- Izračunaćemo mere za svaki „prefiks“: najbolji 1, najboljih 2, najboljih 3, najboljih 4 itd. pogodaka

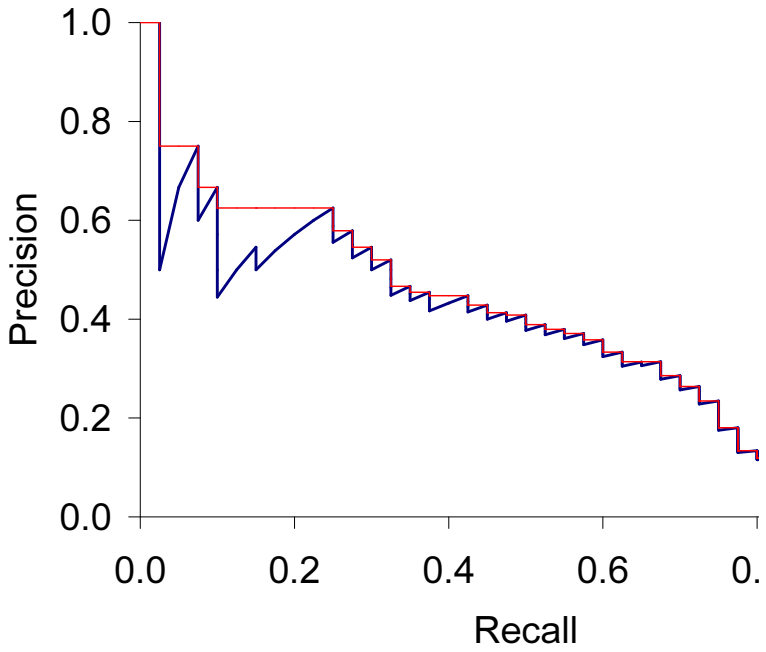
Precision/recall kriva

- Preciznost/povrat/F su mere nerangiranih skupova.
- Lako ih možemo pretvoriti u mere rangiranih lista.
- Izračunaćemo mere za svaki „prefiks“: najbolji 1, najboljih 2, najboljih 3, najboljih 4 itd. pogodaka
- Izračunavanje na ovaj način za preciznost i povrat daje precision/recall krivu.

Precision/recall kriva

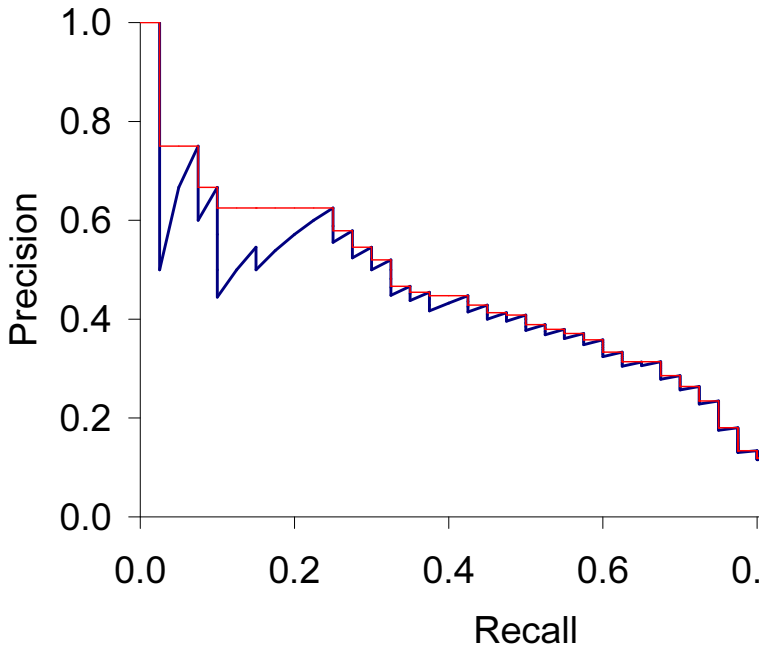


Precision/recall kriva



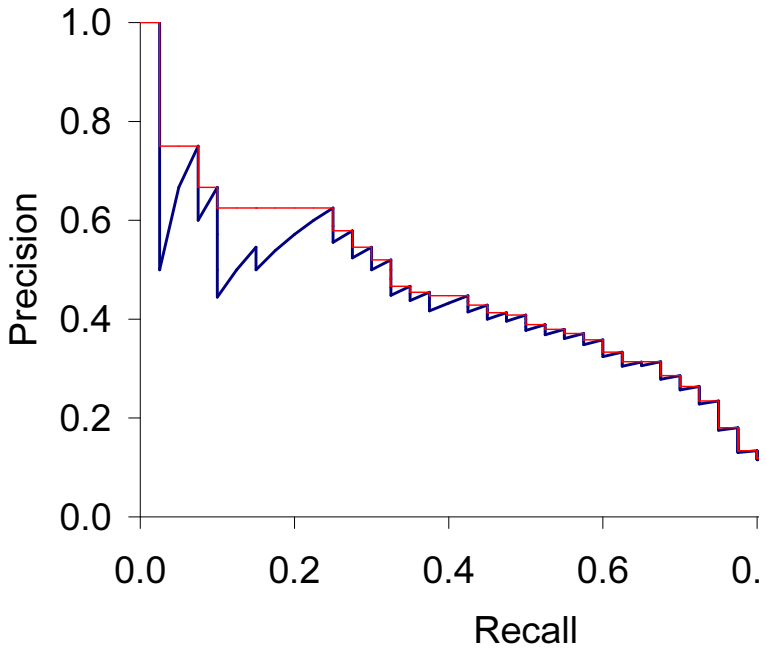
- Svaka tačka odgovara rezultatu za najboljih k rangiranih pogodaka ($k = 1, 2, 3, 4, \ldots$).

Precision/recall kriva



- Svaka tačka odgovara rezultatu za najboljih k rangiranih pogodaka ($k = 1, 2, 3, 4, \ldots$).
- Interpolacija (crveno): Uzmi maksimum svih budućih tačaka

Precision/recall kriva



- Svaka tačka odgovara rezultatu za najboljih k rangiranih pogodaka ($k = 1, 2, 3, 4, \ldots$).
- Interpolacija (crveno): Uzmi maksimum svih budućih tačaka
- Razlog za interpolaciju: Korisnik će hteti da pregleda još pogodaka ako se preciznost i povrat popravljaju

Interpolirana prosečna preciznost u 11 tačaka

Povrat	Interpolirana Preciznost
0.0	1.00
0.1	0.67
0.2	0.63
0.3	0.55
0.4	0.45
0.5	0.41
0.6	0.36
0.7	0.29
0.8	0.13
0.9	0.10
1.0	0.08

Interpolirana prosečna preciznost u 11 tačaka

Povrat	Interpolirana Preciznost
0.0	1.00
0.1	0.67
0.2	0.63
0.3	0.55
0.4	0.45
0.5	0.41
0.6	0.36
0.7	0.29
0.8	0.13
0.9	0.10
1.0	0.08

prosek za 11 tačaka: $\approx$ 0.425

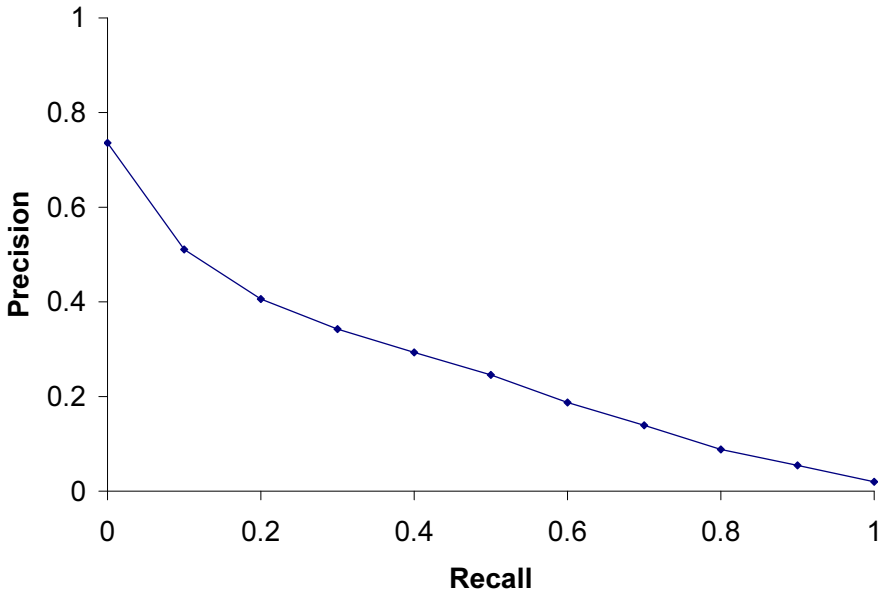
Interpolirana prosečna preciznost u 11 tačaka

Povrat	Interpolirana Preciznost
0.0	1.00
0.1	0.67
0.2	0.63
0.3	0.55
0.4	0.45
0.5	0.41
0.6	0.36
0.7	0.29
0.8	0.13
0.9	0.10
1.0	0.08

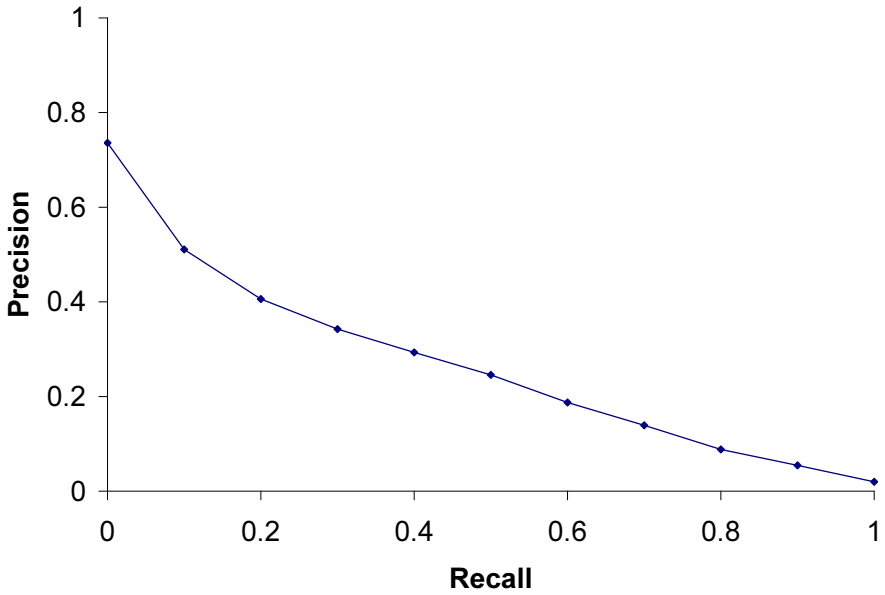
prosek za 11 tačaka: $\approx$ 0.425

Preciznost u 0.0 je > 0?

Uprosečeni precision/recall grafikon za 11 tačaka

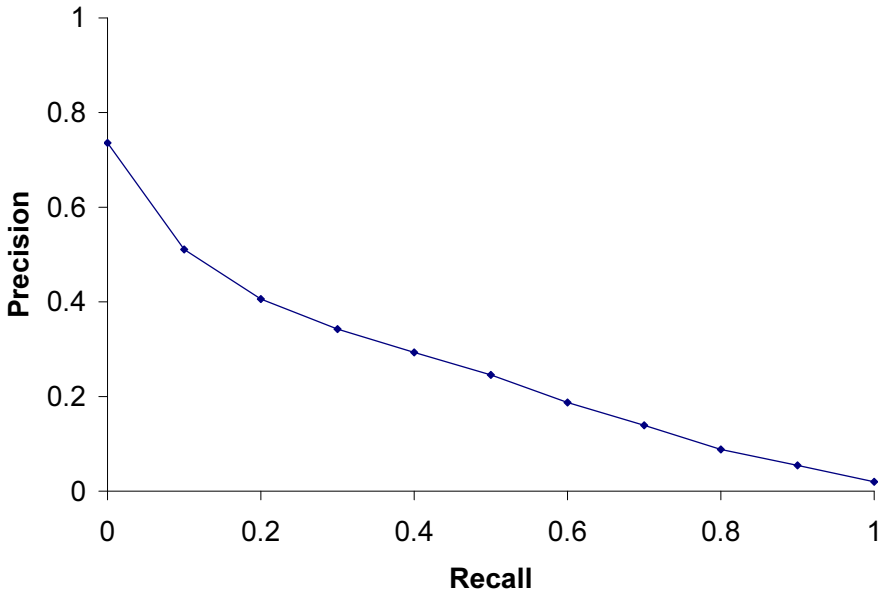


Uprosečeni precision/recall grafikon za 11 tačaka



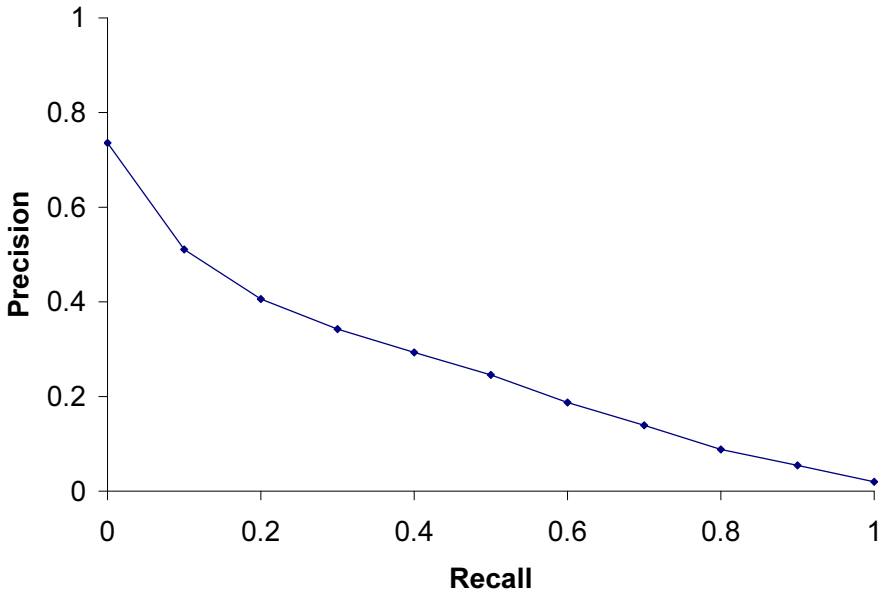
- Izračunaj interpoliranu preciznost za nivoe povrata 0.0, 0.1, 0.2, ...

Uprosečeni precision/recall grafikon za 11 tačaka



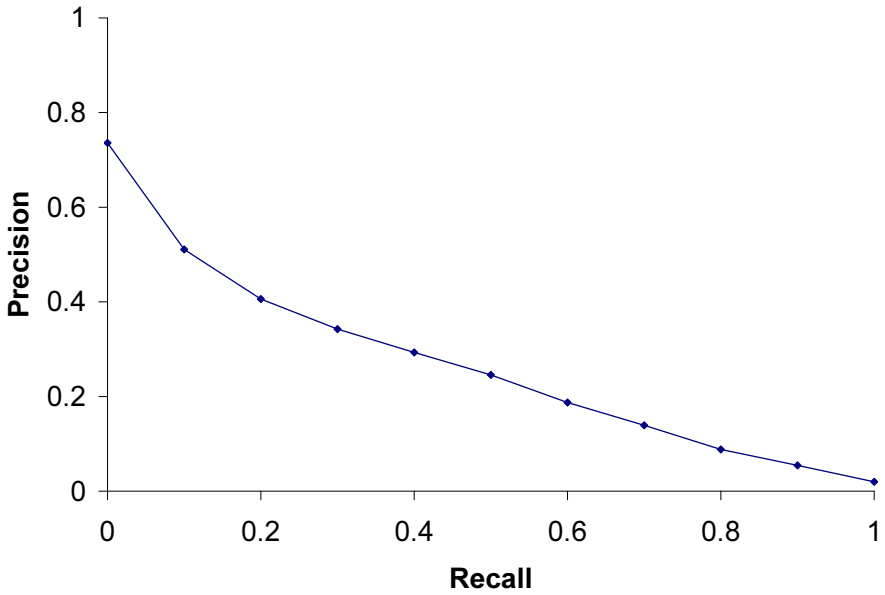
- Izračunaj interpoliranu preciznost za nivoe povrata 0.0, 0.1, 0.2, ...
- Uradi ovo za svaki od test-upita

Uprosečeni precision/recall grafikon za 11 tačaka



- Izračunaj interpoliranu preciznost za nivoe povrata 0.0, 0.1, 0.2, ...
- Uradi ovo za svaki od test-upita
- Uproseči dobijene vrednosti

Uprosečeni precision/recall grafikon za 11 tačaka



- Izračunaj interpoliranu preciznost za nivoe povrata 0.0, 0.1, 0.2, ...
- Uradi ovo za svaki od test-upita
- Uproseči dobijene vrednosti
- Ovo je mera performansi za sve nivoe povrata

Varijansa mere precision/recall

- Za test kolekciju uobičajeno je da sistem radi loše za neke informacione potrebe (npr. $P = 0.2$ za $R = 0.1$) a odlično za neke druge (npr. $P = 0.95$ za $R = 0.1$)

Varijansa mere precision/recall

- Za test kolekciju uobičajeno je da sistem radi loše za neke informacione potrebe (npr. $P = 0.2$ za $R = 0.1$) a odlično za neke druge (npr. $P = 0.95$ za $R = 0.1$)
- Obično je varijansa sistema za više upita mnogo veća nego varijansa različitih sistema za isti upit

Varijansa mere precision/recall

- Za test kolekciju uobičajeno je da sistem radi loše za neke informacione potrebe (npr. $P = 0.2$ za $R = 0.1$) a odlično za neke druge (npr. $P = 0.95$ za $R = 0.1$)
- Obično je varijansa sistema za više upita mnogo veća nego varijansa različitih sistema za isti upit
- Dakle, postoje jednostavne i složene informacione potrebe

Merenje kod velikih pretraživača

- Povrat se teško meri na webu

Merenje kod velikih pretraživača

- Povrat se teško meri na webu
- Pretraživači obično koriste preciznost za najboljih k, npr. $k = 10$ ...

Merenje kod velikih pretraživača

- Povrat se teško meri na webu
- Pretraživači obično koriste preciznost za najboljih k, npr. $k = 10$ ...
- ...ili mere koje više vrednuju da je prvi podogak bolji nego deseti

Merenje kod velikih pretraživača

- Povrat se teško meri na webu
- Pretraživači obično koriste preciznost za najboljih k, npr. $k = 10$ ...
- ...ili mere koje više vrednuju da je prvi podogak bolji nego deseti
- Takođe se koriste mere koje nisu zasnovane na relevantnosti

Merenje kod velikih pretraživača

- Povrat se teško meri na webu
- Pretraživači obično koriste preciznost za najboljih k, npr. $k = 10$ ...
- ...ili mere koje više vrednuju da je prvi podogak bolji nego deseti
- Takođe se koriste mere koje nisu zasnovane na relevantnosti
  - Primer 1: clickthrough za prvi pogodak

Merenje kod velikih pretraživača

- Povrat se teško meri na webu
- Pretraživači obično koriste preciznost za najboljih k, npr. $k = 10$ ...
- ...ili mere koje više vrednuju da je prvi podogak bolji nego deseti
- Takođe se koriste mere koje nisu zasnovane na relevantnosti
  - Primer 1: clickthrough za prvi pogodak
  - Nije vrlo pouzdana ako se posmatra jedan clickthrough (korisnik može da odluči da je dokument nerelevantan nakon uvida u rezime) ...

Merenje kod velikih pretraživača

- Povrat se teško meri na webu
- Pretraživači obično koriste preciznost za najboljih k, npr. $k = 10$ ...
- ...ili mere koje više vrednuju da je prvi podogak bolji nego deseti
- Takođe se koriste mere koje nisu zasnovane na relevantnosti
  - Primer 1: clickthrough za prvi pogodak
  - Nije vrlo pouzdana ako se posmatra jedan clickthrough (korisnik može da odluči da je dokument nerelevantan nakon uvida u rezime) ...
  - ...ali je prilično pouzdana u proseku za veliki broj korisnika.

Merenje kod velikih pretraživača

- Povrat se teško meri na webu
- Pretraživači obično koriste preciznost za najboljih k, npr. $k = 10$ ...
- ...ili mere koje više vrednuju da je prvi podogak bolji nego deseti
- Takođe se koriste mere koje nisu zasnovane na relevantnosti
  - Primer 1: clickthrough za prvi pogodak
  - Nije vrlo pouzdana ako se posmatra jedan clickthrough (korisnik može da odluči da je dokument nerelevantan nakon uvida u rezime) ...
  - ...ali je prilično pouzdana u proseku za veliki broj korisnika.
  - Primer 2: Laboratorijske studije ponašanja korisnika

Merenje kod velikih pretraživača

- Povrat se teško meri na webu
- Pretraživači obično koriste preciznost za najboljih k, npr. $k = 10$ ...
- ...ili mere koje više vrednuju da je prvi podogak bolji nego deseti
- Takođe se koriste mere koje nisu zasnovane na relevantnosti
  - Primer 1: clickthrough za prvi pogodak
  - Nije vrlo pouzdana ako se posmatra jedan clickthrough (korisnik može da odluči da je dokument nerelevantan nakon uvida u rezime) ...
  - ...ali je prilično pouzdana u proseku za veliki broj korisnika.
  - Primer 2: Laboratorijske studije ponašanja korisnika
  - Primer 3: A/B testiranje

A/B testiranje

- Cilj: Testiranje jednog unapređenja

A/B testiranje

- Cilj: Testiranje jednog unapređenja
- Uslov: Imamo veliki pretraživač u pogonu

A/B testiranje

- Cilj: Testiranje jednog unapređenja
- Uslov: Imamo veliki pretraživač u pogonu
- Neka većina korisnika pristupa staroj verziji sistema

A/B testiranje

- Cilj: Testiranje jednog unapređenja
- Uslov: Imamo veliki pretraživač u pogonu
- Neka većina korisnika pristupa staroj verziji sistema
- Skrenućemo mali deo saobraćaja (recimo 1%) na novu verziju sistema koja ima unapređenje

A/B testiranje

- Cilj: Testiranje jednog unapređenja
- Uslov: Imamo veliki pretraživač u pogonu
- Neka većina korisnika pristupa staroj verziji sistema
- Skrenućemo mali deo saobraćaja (recimo 1%) na novu verziju sistema koja ima unapređenje
- Vrednovanje pomoću „automatske“ mere, npr. clickthrough za prvi pogodak

A/B testiranje

- Cilj: Testiranje jednog unapređenja
- Uslov: Imamo veliki pretraživač u pogonu
- Neka većina korisnika pristupa staroj verziji sistema
- Skrenućemo mali deo saobraćaja (recimo 1%) na novu verziju sistema koja ima unapređenje
- Vrednovanje pomoću „automatske“ mere, npr. clickthrough za prvi pogodak
- Sada možemo direktno videti da li unapređenje povećava zadovoljstvo korisnika

A/B testiranje

- Cilj: Testiranje jednog unapređenja
- Uslov: Imamo veliki pretraživač u pogonu
- Neka većina korisnika pristupa staroj verziji sistema
- Skrenućemo mali deo saobraćaja (recimo 1%) na novu verziju sistema koja ima unapređenje
- Vrednovanje pomoću „automatske“ mere, npr. clickthrough za prvi pogodak
- Sada možemo direktno videti da li unapređenje povećava zadovoljstvo korisnika
- Verovatno metodologija kojoj veliki pretraživači najviše veruju

A/B testiranje

- Cilj: Testiranje jednog unapređenja
- Uslov: Imamo veliki pretraživač u pogonu
- Neka većina korisnika pristupa staroj verziji sistema
- Skrenućemo mali deo saobraćaja (recimo 1%) na novu verziju sistema koja ima unapređenje
- Vrednovanje pomoću „automatske“ mere, npr. clickthrough za prvi pogodak
- Sada možemo direktno videti da li unapređenje povećava zadovoljstvo korisnika
- Verovatno metodologija kojoj veliki pretraživači najviše veruju
- Varijanta: dati korisnicima mogućnost da sami izaberu staru ili novu verziju sistema

Šta je potrebno za evaluaciju baziranu na benchmark-u

- Kolekcija dokumenata

Šta je potrebno za evaluaciju baziranu na benchmark-u

- Kolekcija dokumenata
  - dokumenti moraju reprezentovati dokumente koje očekujemo da imamo i u stvarnom slučaju

Šta je potrebno za evaluaciju baziranu na benchmark-u

- Kolekcija dokumenata
  - dokumenti moraju reprezentovati dokumente koje očekujemo da imamo i u stvarnom slučaju
- Kolekcija informacionih potreba

Šta je potrebno za evaluaciju baziranu na benchmark-u

- Kolekcija dokumenata
  - dokumenti moraju reprezentovati dokumente koje očekujemo da imamo i u stvarnom slučaju
- Kolekcija informacionih potreba
  - . . . koje ćemo često neispravno nazivati upitima

Šta je potrebno za evaluaciju baziranu na benchmark-u

- Kolekcija dokumenata
  - dokumenti moraju reprezentovati dokumente koje očekujemo da imamo i u stvarnom slučaju
- Kolekcija informacionih potreba
  - . . . koje ćemo često neispravno nazivati upitima
  - informacione potrebe moraju reprezentovati one koje očekujemo i u stvarnom slučaju

Šta je potrebno za evaluaciju baziranu na benchmark-u

- Kolekcija dokumenata
  - dokumenti moraju reprezentovati dokumente koje očekujemo da imamo i u stvarnom slučaju
- Kolekcija informacionih potreba
  - . . . koje ćemo često neispravno nazivati upitima
  - informacione potrebe moraju reprezentovati one koje očekujemo i u stvarnom slučaju
- Čovekove ocene relevantnosti

Šta je potrebno za evaluaciju baziranu na benchmark-u

- Kolekcija dokumenata
  - dokumenti moraju reprezentovati dokumente koje očekujemo da imamo i u stvarnom slučaju
- Kolekcija informacionih potreba
  - . . . koje ćemo često neispravno nazivati upitima
  - informacione potrebe moraju reprezentovati one koje očekujemo i u stvarnom slučaju
- Čovekove ocene relevantnosti
  - moramo angažovati ocenjivače za ovaj posao

Šta je potrebno za evaluaciju baziranu na benchmark-u

- Kolekcija dokumenata
  - dokumenti moraju reprezentovati dokumente koje očekujemo da imamo i u stvarnom slučaju
- Kolekcija informacionih potreba
  - . . . koje ćemo često neispravno nazivati upitima
  - informacione potrebe moraju reprezentovati one koje očekujemo i u stvarnom slučaju
- Čovekove ocene relevantnosti
  - moramo angažovati ocenjivače za ovaj posao
  - skupo, troši puno vremena

Šta je potrebno za evaluaciju baziranu na benchmark-u

- Kolekcija dokumenata
  - dokumenti moraju reprezentovati dokumente koje očekujemo da imamo i u stvarnom slučaju
- Kolekcija informacionih potreba
  - . . . koje ćemo često neispravno nazivati upitima
  - informacione potrebe moraju reprezentovati one koje očekujemo i u stvarnom slučaju
- Čovekove ocene relevantnosti
  - moramo angažovati ocenjivače za ovaj posao
  - skupo, troši puno vremena
  - ocenjivači moraju reprezentovati one koje očekujemo i u stvarnom slučaju

Šta je potrebno za evaluaciju baziranu na benchmark-u

- Kolekcija dokumenata
  - dokumenti moraju reprezentovati dokumente koje očekujemo da imamo i u stvarnom slučaju
- Kolekcija informacionih potreba
  - . . . koje ćemo često neispravno nazivati upitima
  - informacione potrebe moraju reprezentovati one koje očekujemo i u stvarnom slučaju
- Čovekove ocene relevantnosti
  - moramo angažovati ocenjivače za ovaj posao
  - skupo, troši puno vremena
  - ocenjivači moraju reprezentovati one koje očekujemo i u stvarnom slučaju
  - Ocene relevantnosti su korisne samo ako su konzistentne.

Šta je potrebno za evaluaciju baziranu na benchmark-u

- Kolekcija dokumenata
  - dokumenti moraju reprezentovati dokumente koje očekujemo da imamo i u stvarnom slučaju
- Kolekcija informacionih potreba
  - . . . koje ćemo često neispravno nazivati upitima
  - informacione potrebe moraju reprezentovati one koje očekujemo i u stvarnom slučaju
- Čovekove ocene relevantnosti
  - moramo angažovati ocenjivače za ovaj posao
  - skupo, troši puno vremena
  - ocenjivači moraju reprezentovati one koje očekujemo i u stvarnom slučaju
  - Ocene relevantnosti su korisne samo ako su konzistentne.
  - Kako možemo meriti konzistentnost među ocenjivačima?

Šta je potrebno za evaluaciju baziranu na benchmark-u

- Kolekcija dokumenata
  - dokumenti moraju reprezentovati dokumente koje očekujemo da imamo i u stvarnom slučaju
- Kolekcija informacionih potreba
  - . . . koje ćemo često neispravno nazivati upitima
  - informacione potrebe moraju reprezentovati one koje očekujemo i u stvarnom slučaju
- Čovekove ocene relevantnosti
  - moramo angažovati ocenjivače za ovaj posao
  - skupo, troši puno vremena
  - ocenjivači moraju reprezentovati one koje očekujemo i u stvarnom slučaju
  - Ocene relevantnosti su korisne samo ako su konzistentne.
  - Kako možemo meriti konzistentnost među ocenjivačima?

Šta je potrebno za evaluaciju baziranu na benchmark-u

- Kolekcija dokumenata
  - dokumenti moraju reprezentovati dokumente koje očekujemo da imamo i u stvarnom slučaju
- Kolekcija informacionih potreba
  - . . . koje ćemo često neispravno nazivati upitima
  - informacione potrebe moraju reprezentovati one koje očekujemo i u stvarnom slučaju
- Čovekove ocene relevantnosti
  - moramo angažovati ocenjivače za ovaj posao
  - skupo, troši puno vremena
  - ocenjivači moraju reprezentovati one koje očekujemo i u stvarnom slučaju
  - Ocene relevantnosti su korisne samo ako su konzistentne.
  - Kako možemo meriti konzistentnost među ocenjivačima? Kapa mera

κ: kapa mera

- Kapa je mera koliko se međusobno ocenjivači slažu

κ: kapa mera

- Kapa je mera koliko se međusobno ocenjivači slažu
- Dizajnirana za kategorične ocene

κ: kapa mera

- Kapa je mera koliko se međusobno ocenjivači slažu
- Dizajnirana za kategorične ocene
- $P(A) =$ koji deo od ukupnog broja slučajeva se ocenjivači slažu

κ: kapa mera

- Kapa je mera koliko se međusobno ocenjivači slažu
- Dizajnirana za kategorične ocene
- $P(A) =$ koji deo od ukupnog broja slučajeva se ocenjivači slažu
- $P(E) =$ koji deo slaganja bismo dobili slučajno

κ: kapa mera

- Kapa je mera koliko se međusobno ocenjivači slažu
- Dizajnirana za kategorične ocene
- $P(A) =$ koji deo od ukupnog broja slučajeva se ocenjivači slažu
- $P(E) =$ koji deo slaganja bismo dobili slučajno
-

$$\kappa = \frac{P(A) - P(E)}{1 - P(E)}$$

κ: kapa mera

- Kapa je mera koliko se međusobno ocenjivači slažu
- Dizajnirana za kategorične ocene
- $P(A) =$ koji deo od ukupnog broja slučajeva se ocenjivači slažu
- $P(E) =$ koji deo slaganja bismo dobili slučajno
- $$\kappa = \frac{P(A) - P(E)}{1 - P(E)}$$
- $\kappa =?$ za (i) slučajno slaganje (ii) totalno slaganje

κ: kapa mera

- Vrednosti $\kappa \in [2/3, 1.0]$ se smatraju prihvatljivim

κ: kapa mera

- Vrednosti $\kappa \in [2/3, 1.0]$ se smatraju prihvatljivim
- Sa manjim vrednostima: potreban je redizajn metodologije ocenjivanja itd.

Izračunavanje κ statistike

		Ocenj. 2 relevantnost		
		Da	Ne	Total
Ocenj. 1 relevantnost	Da	300	20	320
	Ne	10	70	80
	Total	310	90	400

Odnos puta kada su se ocenjivači složili
$P(A) = (300 + 70)/400 = 370/400 = 0.925$
Pooled marginals
$P(nerelevantan) = (80 + 90)/(400 + 400) = 170/800 = 0.2125$
$P(relevantan) = (320 + 310)/(400 + 400) = 630/800 = 0.7878$
Verovatnoća da su se slučajno složili $P(E) =$
$P(nerelevantan)^2 + P(relevantan)^2 = 0.2125^2 + 0.7878^2 = 0.665$
Kapa mera
$\kappa = (P(A) - P(E))/(1 - P(E)) = (0.925 - 0.665)/(1 - 0.665) = 0.776$

Izračunavanje κ statistike

		Ocenj. 2 relevantnost		
		Da	Ne	Total
Ocenj. 1 relevantnost	Da	300	20	320
	Ne	10	70	80
	Total	310	90	400

Odnos puta kada su se ocenjivači složili
$P(A) = (300 + 70)/400 = 370/400 = 0.925$
Pooled marginals
$P(nerelevantan) = (80 + 90)/(400 + 400) = 170/800 = 0.2125$
$P(relevantan) = (320 + 310)/(400 + 400) = 630/800 = 0.7878$
Verovatnoća da su se slučajno složili $P(E) =$
$P(nerelevantan)^2 + P(relevantan)^2 = 0.2125^2 + 0.7878^2 = 0.665$
Kapa mera
$\kappa = (P(A) - P(E))/(1 - P(E)) = (0.925 - 0.665)/(1 - 0.665) = 0.776$
(i dalje prihvatljivo)

Standardni benchmark za relevantnost: Cranfield

- Prvi skup testova za precizno merenje efektivnosti IR sistema

Standardni benchmark za relevantnost: Cranfield

- Prvi skup testova za precizno merenje efektivnosti IR sistema
- Kasne 1950te, UK

Standardni benchmark za relevantnost: Cranfield

- Prvi skup testova za precizno merenje efektivnosti IR sistema
- Kasne 1950te, UK
- 1398 apstrakata iz članaka o aerodinamici, skup od 225 upita, iscrpne ocene relevantnosti za sve parove upit-dokument

Standardni benchmark za relevantnost: Cranfield

- Prvi skup testova za precizno merenje efektivnosti IR sistema
- Kasne 1950te, UK
- 1398 apstrakata iz članaka o aerodinamici, skup od 225 upita, iscrpne ocene relevantnosti za sve parove upit-dokument
- Za današnje uslove suviše mali i atipičan uzorak

Standardni benchmark za relevantnost: TREC

- TREC = Text Retrieval Conference (TREC)

Standardni benchmark za relevantnost: TREC

- TREC = Text Retrieval Conference (TREC)
- Organizuje U.S. National Institute of Standards and Technology (NIST)

Standardni benchmark za relevantnost: TREC

- TREC = Text Retrieval Conference (TREC)
- Organizuje U.S. National Institute of Standards and Technology (NIST)
- TREC je skup različitih bechmarka za relevantnost

Standardni benchmark za relevantnost: TREC

- TREC = Text Retrieval Conference (TREC)
- Organizuje U.S. National Institute of Standards and Technology (NIST)
- TREC je skup različitih bechmarka za relevantnost
- Najpoznatiji: TREC Ad Hoc, korišćen za prvih 8 TREC sastanaka između 1992 i 1999

Standardni benchmark za relevantnost: TREC

- TREC = Text Retrieval Conference (TREC)
- Organizuje U.S. National Institute of Standards and Technology (NIST)
- TREC je skup različitih bechmarka za relevantnost
- Najpoznatiji: TREC Ad Hoc, korišćen za prvih 8 TREC sastanaka između 1992 i 1999
- 1.89 milion dokumenata, uglavnom novinskih članaka, 450 informacionih potreba

Standardni benchmark za relevantnost: TREC

- TREC = Text Retrieval Conference (TREC)
- Organizuje U.S. National Institute of Standards and Technology (NIST)
- TREC je skup različitih bechmarka za relevantnost
- Najpoznatiji: TREC Ad Hoc, korišćen za prvih 8 TREC sastanaka između 1992 i 1999
- 1.89 milion dokumenata, uglavnom novinskih članaka, 450 informacionih potreba
- Nema iscrpnih ocena relevantnosti – previše skupo

Standardni benchmark za relevantnost: TREC

- TREC = Text Retrieval Conference (TREC)
- Organizuje U.S. National Institute of Standards and Technology (NIST)
- TREC je skup različitih bechmarka za relevantnost
- Najpoznatiji: TREC Ad Hoc, korišćen za prvih 8 TREC sastanaka između 1992 i 1999
- 1.89 milion dokumenata, uglavnom novinskih članaka, 450 informacionih potreba
- Nema iscrpnih ocena relevantnosti – previše skupo
- Umesto toga, ocene NIST-ovih ocenjivača postoje samo za dokumente koji su bili među prvih k koje je vratio jedan od sistema u TREC testu

Drugi standardni benchmarci za relevantnost

- GOV2

Drugi standardni benchmarci za relevantnost

- GOV2
  - još jedna TREC/NIST kolekcija

Drugi standardni benchmarci za relevantnost

- GOV2
  - još jedna TREC/NIST kolekcija
  - 25 miliona web strana

Drugi standardni benchmarci za relevantnost

- GOV2
  - još jedna TREC/NIST kolekcija
  - 25 miliona web strana
  - najveća kolekcija koja je lako dostupna

Drugi standardni benchmarci za relevantnost

- GOV2
  - još jedna TREC/NIST kolekcija
  - 25 miliona web strana
  - najveća kolekcija koja je lako dostupna
  - ali i dalje 3 reda veličine manja od Google/Yahoo indeksa

Drugi standardni benchmarci za relevantnost

- GOV2
  - još jedna TREC/NIST kolekcija
  - 25 miliona web strana
  - najveća kolekcija koja je lako dostupna
  - ali i dalje 3 reda veličine manja od Google/Yahoo indeksa
- NTCIR

Drugi standardni benchmarci za relevantnost

- GOV2
  - još jedna TREC/NIST kolekcija
  - 25 miliona web strana
  - najveća kolekcija koja je lako dostupna
  - ali i dalje 3 reda veličine manja od Google/Yahoo indeksa
- NTCIR
  - IR za dalekoistočne jezike

Drugi standardni benchmarci za relevantnost

- GOV2
  - još jedna TREC/NIST kolekcija
  - 25 miliona web strana
  - najveća kolekcija koja je lako dostupna
  - ali i dalje 3 reda veličine manja od Google/Yahoo indeksa
- NTCIR
  - IR za dalekoistočne jezike
- Cross Language Evaluation Forum (CLEF)

Drugi standardni benchmarci za relevantnost

- GOV2
  - još jedna TREC/NIST kolekcija
  - 25 miliona web strana
  - najveća kolekcija koja je lako dostupna
  - ali i dalje 3 reda veličine manja od Google/Yahoo indeksa
- NTCIR
  - IR za dalekoistočne jezike
- Cross Language Evaluation Forum (CLEF)
  - fokusiran na evropske jezike i cross-language pretraživanje

Drugi standardni benchmarci za relevantnost

- GOV2
  - još jedna TREC/NIST kolekcija
  - 25 miliona web strana
  - najveća kolekcija koja je lako dostupna
  - ali i dalje 3 reda veličine manja od Google/Yahoo indeksa
- NTCIR
  - IR za dalekoistočne jezike
- Cross Language Evaluation Forum (CLEF)
  - fokusiran na evropske jezike i cross-language pretraživanje
- Mnogi drugi

INEX

- Benchmark za ocenu performansi pretraživanja XML-a
  - analogno TREC-u
- Sastoji se iz
  - skupa XML dokumenata
  - kolekcije pretraživačkih zadataka

INEX

- Svaki sistem indeksira dokumente
- Autori sistema zapisuju pretraživačke zadatke kao upite
  - koristeći upitni jezik koji sistem razume
- U odgovoru, sistem vraća elemente unutar dokumenata (ne cele dokumente)
- Rangira pronađene elemente

INEX ocena

- Za svaki upit, svaki pronađeni element se ocenjuje po dva kriterijuma
  - relevantnost: koliko je relevantan element
  - pokrivanje: da li je element suviše uzak ili suviše širok
  - npr. za upit koji traži definiciju Furijeove transformacije: da li ćemo dobiti samo jednačinu (suviše uzak), celo poglavlje (suviše širok), ili tekst definicije
- Ove ocene se koriste za izračunavanje precision/recall mera

INEX kolekcija

- 12.107 publikacija iz IEEE Computer Society
- 494 megabajta
- Prosečan članak: 1532 XML čvora
  - prosečna dubina: 6.9

INEX teme

- Svaka tema (topic) predstavlja informacionu potrebu
- Dve vrste tema
  - content only (CO): free-text upiti
  - content and structure (CAS): eksplicitna ograničenja na strukturu dokumenta, tj. sadržavanje elemenata

INEX ocene

- Svaki pretraživač formuliše temu kao upit
  - npr. korišćenjem ključnih reči u temi
- Pretraživač pronalazi odgovarajuće elemente i rangira ih
- Eksperti-ocenjivači dodeljuju svakom pronađenom elementu ocene za relevantnost i pokrivanje

INEX ocene

- Relevantnost se ocenjuje na skali od 0 (irelevantno) do 3 (vrlo relevantno)
- Pokrivanje se ocenjuje na skali od četiri ocene:
  - **No coverage**: tema ne odgovara ničemu u elementu
  - **Too large**: tema je manji deo pronađenog elementa
  - **Too small**: element je premali da pokrije temu
  - **Exact**
- Svaki pronađeni element ima ocenu iz skupa $\{0, 1, 2, 3\} \times \{N, S, L, E\}$

Kombinovanje ocena

$$f_{strict}(rel, cov) = \begin{cases} 1 & (rel, cov) = 3E \\ 0 & inače \end{cases}$$

$$f_{generalized}(rel, cov) = \begin{cases} 1.00 & (rel, cov) = 3E \\ 0.75 & (rel, cov) \in \{2E, 3L, 3S\} \\ 0.50 & (rel, cov) \in \{1E, 2L, 2S\} \\ 0.25 & (rel, cov) \in \{1S, 1L\} \\ 0.00 & (rel, cov) = 0N \end{cases}$$

f-vrednosti

- Skalarna mera kvaliteta pronađenog elementa
- Mogu se izračunati f-vrednosti za različite brojeve pronađenih elemenata: 10, 20, itd.
  - sredstvo za poređenje pretraživača