

UNIVERZITET SINGIDUNUM
DEPARTMAN ZA POSLEDIPLOMSKE STUDIJE

MASTER RAD

ANALIZA SISTEMA ZA PRETRAGU SLIKA I TEKSTA

Mentor:

Prof. dr Milan Milosavljević

Student:

Aleksandar Živković
Br. indeksa 410449/2014

Beograd, 2017. godine

Sadržaj

<u>Apstrakt</u>	<u>- 5 -</u>
<u>Uvod</u>	<u>- 6 -</u>
<u>1 Pronalaženje informacija</u>	<u>- 10 -</u>
<u>2 Modeli sistema za pronalaženje informacija</u>	<u>- 10 -</u>
2.1 Model vektorskog prostora	- 11 -
2.2 Modeli verovatnoće	- 13 -
2.3 Model mreže zaključivanja	- 15 -
<u>3 Implementacija.....</u>	<u>- 16 -</u>
3.1 BM25 rankirajuća funkcija	- 17 -
3.2 PageRank algoritam	- 18 -
3.3 HITS algoritam	- 19 -
<u>4 Ocenjivanje sistema</u>	<u>- 20 -</u>
<u>5 Modifikacija upita.....</u>	<u>- 21 -</u>
5.1 Upiti u obliku pitanja	- 22 -

6	<u>Pretraživanje slika bazirano na tekstu</u>	<u>- 23 -</u>
6.1	Selekcija obima pretrage.....	- 24 –
6.2	Konstrukcija indeksa.....	- 25 –
6.3	Frekvenca termina – frekvenca obrnutih dokumenata.....	- 26 –
6.4	Indeksiranje obrnutih dokumenata.....	- 27 –
7	<u>Pretraživanje slika bazirano na sadržaju</u>	<u>- 28 –</u>
7.1	Predstavljanje slika i teksta	- 29–
7.1.1	Histogrami boja.....	- 30 –
7.1.2	Tamura odlike tekstura	- 33 –
7.1.3	Raspored histograma.....	- 34 –
7.1.4	Lokalne odlike	- 35 –
7.1.5	Lokalni histogrami boja	- 36 –
7.1.6	GIST.....	- 36 –

8	Povratne informacije o relevantnosti	- 37-
8.1	Metode razdaljine.....	- 38 -
8.1.1	Euklidova razdaljina	- 39 -
8.1.2	L_1	- 40 -
8.1.3	L_∞	- 40 -
8.1.4	Dzensen-Šanon	- 41 -
8.1.5	Kulbak-Leibler	- 41 -
8.1.6	Distorzija.....	- 41 -
8.1.7	Razdaljina vizuelnih reči.....	- 42 -
8.2	Metode ocenjivanja.....	- 43 -
9	Zaključna razmatranja	- 46 -
	<u>Literatura.....</u>	- 47 -

Apstrakt

U današnjem svakodnevnom životu, važnost arhivisanja i pronalaženja informacija postaje sve više bitnija. U ovom radu objasnićemo neke ključne metodologije i tehnike, kao i evoluciju oblasti pretraživanja informacija. Zaključujemo da je istraživanje u ovoj oblasti od ključnog značaja za brzu i efikasnu dostupnost informacijama što je značajno da skoro svaku ljudku delatnost u modernom svetu.

Ključne reči: veštačka inteligencija, sistem za pronalaženje informacija, pretraživač informacija, pretraživač slika, upit

UVOD

Još od prastarih vremena ljudi su pokušavali da razumeju kako ljudi misle, to jest kako sagledavamo, razumemo, predviđamo i manipulišemo svet koji je mnogo veći i kompleksniji od nas.

Oblast *Veštačke Inteligencije* (VI) se bavi ne samo razumevanjem inteligencije već i konstrukcijom inteligentih sistema. Veštačka Inteligencija je jedna od najnovijih polja u nauci i inženjerstvu, prva istraživanja su pokrenuta posle Drugog Svetskog Rata a samo ime je dizajnirano 1956 godine. Trenutno VI se bavi raznim podpoljima, kako generalnim (učenje i percepcija) tako i specifičnim kao što su igranje šaha, dokazivanje matematičkih teorema, pisanje pozije, voznja automobila i dijagnoza bolesti. VI je univerzalna oblast, bitna je za svaki intelektualni zadatak.

Sama definicija VI nije univerzalno uspostavljena, postoji velika količina njih i sve se zasnivaju na mašinama koje misle i deluju kao ljudi, to jest racionalno. Možda najpoznatija definicija je takozvani *Tjuringov test*, koji je napravio Alan Tjuring 1950. godine.[1] Po njemu čovek ispituje kompjuter ne znajući da li ispituje kompjuter ili drugog čoveka i ako na kraju ne može to da odredi test je položen. Da bi položio ovaj test kompjuter mora da ima sledeće mogućnosti:

- obrađivanje prirodnog jezika – da bi mogao da razgovara sa ispitanikom
- reprezentaciju znanja – da bi mogao da čuva šta zna ili sazna
- automatski razum – da može da koristi sačuvane informacije da odgovara na pitanja i dolazi do novih zaključaka
- mašinsko učenje – da bi mogao da se prilagodi novim situacijama i da primeti i koristi šablone
- kompjuterska vizija – da bi mogao da vidi objekte
- robotika – da bi mogao da manipuliše objektima i da se kreće

Ovih šest disciplina sadrže većinu oblasti iz VI danas i *Tjuringov test* ostaje relevantan i posle 60 godina, ali većina istraživača VI se ne trude da ga polože jer smatraju da je mnogo važnije izučavati bazične principe inteligencije nego duplirati najbolji model (čoveka).

Mnoge druge nauke su doprinele kreiranju i razvoju VI kao što su : filozofija, psihologija, matematika, ekonomija, neuronauka, kompjutersko inženjerstvo, lingvistika.

Iz temelja veštačke inteligencije, specifično obrađivanja prirodnog jezika, nastala je nova naučna oblast zvana pronalaženje informacija.

Čuvanje pisane informacije se može naći i 3000 godina pre Isusa Hrista, kada su Sumeri odvajali specijalna mesta za čuvanje glinenih tabli sa zapisima na klinastom pismu. Čak i Sumeri su shvatili da prava organizacija i pristup arhivama je kritična za efikasno korišćenje informacije. Pravili su specijalne klasifikacije za identifikaciju svake table i njene sadržine.

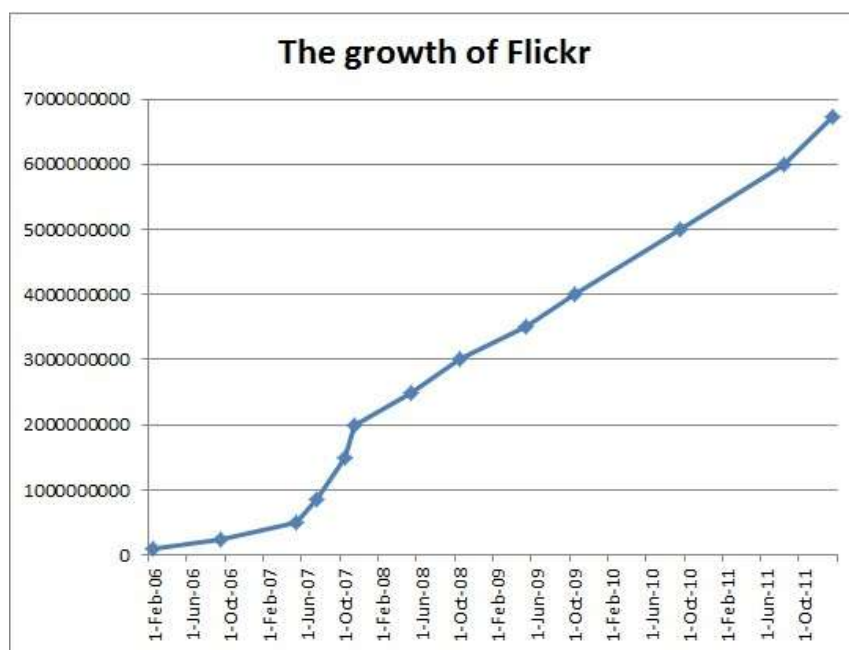
Potreba za skladištenje i pronalaženje informacija je postala sve važnija tokom vekova, naročito sa izumima kao što su papir i presa za štampanje. Uskoro su i računari izumljeni i ljudi su shvatili da mogu da se koriste za skladištenje i mehaničko pronalaženje velike količine informacija. 1965-te godine Vannevar Bush je objavio revolucionarni članak "Kako možemo misliti" koji je rodio ideju automatskog pristupa velikoj količini informacija.[2] 1950-ih godina ova ideja se materijalizovala u konkretne opise kako bi arhive teksta mogle da se automatski pretražuju. Nekoliko radova iz sredine 1950-ih godina su razradili bazičnu ideju pretrage teksta pomoću računara.

Nekoliko glavnih razvitaka u oblasti su se desili 1960-ih godina. Najbitnije je bilo razvije SMART sistema od Gerard Salton-a i njegovih studenata, prvo na Harvard Univerzitetu pa kasnije na Cornell Univerzitetu,[3] i Cranfield procene rađene od Cyril Cleverdon-a i njegove grupe na Cranfield Univerzitetu.[4] Cranfield testovi su razvili metodologiju ocenjivanja sistema za pronalaženje informacija koja se koristi još danas. SMART sistem je dozvolio istraživačima da eksperimentišu sa idejama da bi unapredili kvalitet pretrage. Sistem za eksperimentisanje zajedno da dobrom metodologijom ocenjivanja je doprinelo brzom razvoju u oblasti i mnogim razvićima.

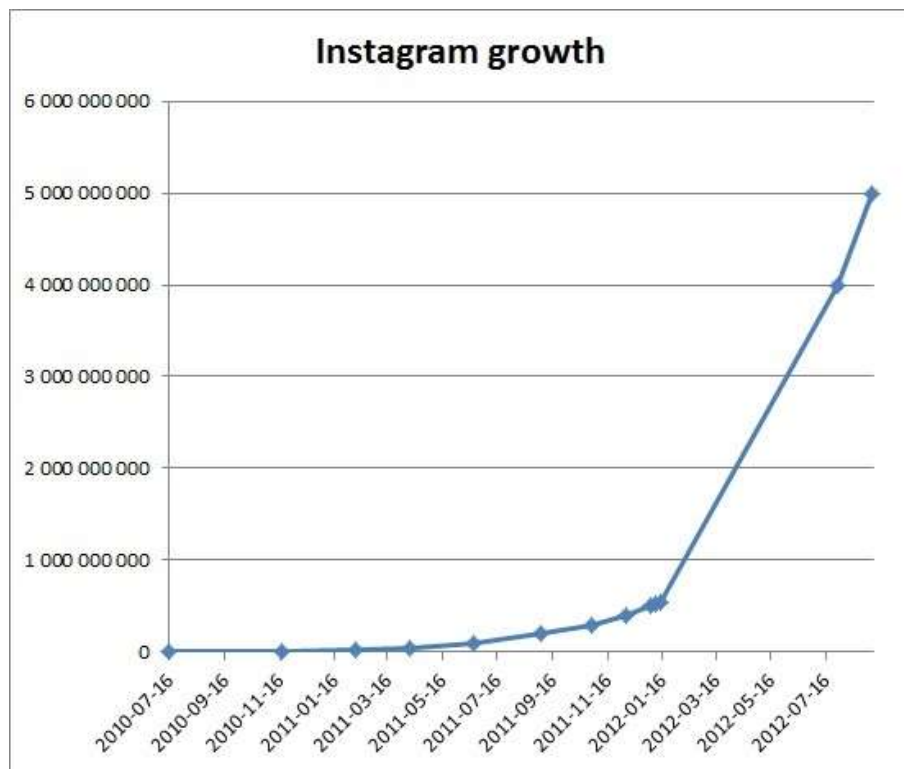
1970-tih i 1980-tih godina su se razvila mnoga unapređenja od ideja 1960-ih. Novi modeli i tehnike su eksperimentalno dokazani da su efikasni na malim kolekcijama teksta (nekoliko hiljada članaka) koji su bili dostupni istraživačima tada, ali pošto nisu imali pristup velikim kolekcijama dokumenata nisu znali da li će biti isto toliko efikasne na većim kolekcijama dokumenata. Ovo se promenilo 1992 godine sa kreiranjem konferencije za pretraživanje teksta (Text Retrieval Conference - TREC).[5]

Sa velikim kolekcijama teksta dostupnim od TREC-a, dosta starih tehnika se modifikovalo i dosta novih se razvilo za efektivnu pretragu velikih kolekcija. Algoritmi razvijeni u ovoj naučnoj oblasti su se prvi koristili za pretragu World Wide Web-a od 1996 do 1998. Web pretraživači su se kasnije razvili da koriste relacione veze dostupne na web-u.

Broj slika na internetu eksponencijalno raste svakog dana zbog faktora kao što su povećana brzina interneta, masovna dostupnost digitalnom kamerama i mobilnim telefonima i socijalne mreže. U modernoj svakodnevnicu je sasvim normalno da svako poseduje bar jednu kameru, kao samu napravu ili u paketu sa mobilnim telefonom. Lakoća samog procesa pravljenja slika u odnosu na stare filmske kamere, kao i opadanje cena memorijskog hardvera je dovela do rasta privatnih kolekcija slika. Dodatno, popularnost socijalnih mreža kao što je Facebook i sajtova za prosleđivanje slika kao što su Flickr i Instagram dodatno podstiču korisnike da prave slike. Zbog ovih razloga tačna klasifikacija i skladištenje slika je bitan problem današnjice. Kada kolekcije slika značajno porastu nije više moguće da jedna osoba sama nađe šta traži, već joj je potrebna pomoć računara.



Slika 1 – Nagli rast korisnika Flickr-a



Slika 2 – Nagli rast korisnika Instagram-a

Kada korisnik traži neku sliku on ima veoma apstraktnu ideju toga što traži, i prevesti to u nešto što računar razume se pokazalo kao veliki problem. Trenutni sistemi za pretraživanje slika pokušavaju da reše ovaj problem praveći pretraživanje slika baziranim na *semantičkim konceptima*. Trenutne tehnike još uvek nisu dovoljne da automatski razumeju semantička značenja slika baziranim na ljudskoj percepciji. Ovaj problem je poznat pod imenom *semantički raskorak*, i uobičajeno se javlja kada se ljudske aktivnosti pokušavaju prevesti u računsku reprezentaciju [20]. Postoje razni načini za pretraživanje slika, ali se najčešće koriste *pretraživanje slika bazirano na tekstu (konceptu)* i *pretraživanje slika bazirano na sadržaju*.

1 PRONALAŽENJE INFORMACIJA

Ovaj rad će se baviti problemom pronalaženja dokumenata koja sadrže informacije relevantne za korisnika. Najbolji primeri za sisteme za pronalaženje informacija su web pretraživači. Korisnik web-a može da ukuca upit kao što je “istorija Srbije” u pretraživač i da vidi sve relevantne stranice. Sistem za pronalaženje informacija se može karakterizovati po:

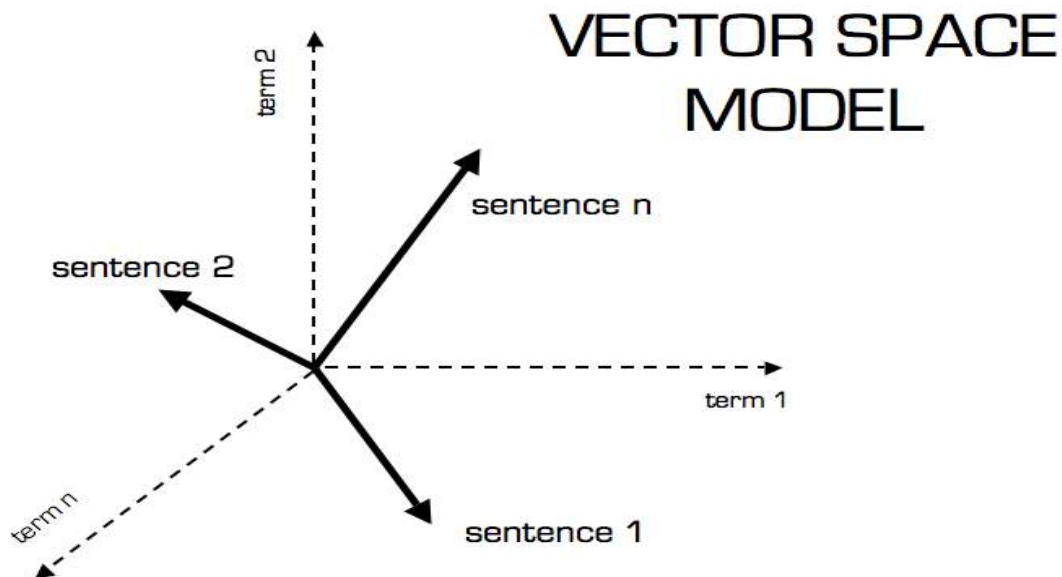
- Definiciji dokumenta – Svaki sistem mora da odluči šta želi da tretira kao dokument: paragraf, stranicu ili više stranica teksta
- Upiti postavljeni u jeziku upita – Upit određuje šta korisnik želi da zna. Jezik upita može biti samo lista reči kao što su [istorija Srbije]; ili može da sadrži frazu reči koje moraju biti jedno pored drugog [“istorija Srbije”]; može da sadrži i logičke operatore kao što su [istorija I Srbije]; i može da sadrži i ne-logičke operatore kao što su [istorija BLIZU Srbije] ili [istorija Srbije stranica:www.srpskaistorija.org].
- Skup rezultata – Ovo je podskup dokumenata koje sistem za pronalaženje informacija smatra relevantnim za upit. Pod relevantnim, smatramo da će verovatno biti od koristi osobi koja je postavila upit, za specifičnu informaciju traženu u upitu.
- Prezentacija skupa rezultata – Ovo može biti ili jednostavno kao rankirana lista titla dokumenata ili komplikovano kao rotirajući tri-dimenzionalni grafikon od skupa rezultata

2 MODELI SISTEMA ZA PRONALAŽENJE INFORMACIJA

Početni sistemi za pronalaženje informacija su radili na modelu *Bulove ključne reči*. Svaka reč u kolekciji dokumenata se tretira kao Bulova vrednost koja je tačna za dokument ako se reč sadrži u tom dokumentu i netačna ako se ne sadrži. Ovaj model ima prednost što je jednostavan za objašnjavanje i implementaciju, ali ima i svojih mana. Nivo relevantnosti dokumenta je jedan bit, tako da nema nikakvog načina da se relevantni dokumenti sortiraju u prezentaciji. Još jedna mana je to što korisnici koji nisu programeri ili matematičari verovatno nisu upoznati sa Bulovom logikom. Ovakvi korisnici će se zbuniti pri korišćenju pretraživača zato što ako žele da znaju vremensku prognozu u Beogradu I Nišu moraju da naprave upit [vreme (Beograd ILI Niš)]. Takodje, pravljenje dobrog upita nekad može biti jako teško i za iskusne korisnike. Zato većina današnjih sistema koriste rankirajuće pronalaženje informacija. Ovi sistemi rankiraju dokumenta po njihovoj proceni o korisnosti dokumanta za korisnikov upit. Većina ovakvih sistema dodeljuje numeričku ocenu za svaki dokument i rankiraju dokumenta po toj oceni. Tri najviše korišćenih modela u istraživanju sistema za pronalaženje informacija su model vektorskog prostora, modeli verovatnoće i model mreža zaključivanja.

2.1 Model vektorskog prostora

U vektorskom modelu tekst je predstavljen od vektora *termina*. [6] Definicija termina ne postoji u modelu, ali termini su uglavnom reči ili fraze. Ako su reči izabrane kao termini, onda je svaka reč u rečniku postaje nezavisna dimenzija u višedimenzionalnom vektorskom prostoru. Bilo koji tekst se onda može predstaviti kao vektor u ovom višedimenzionalnom prostoru. Ako termin pripada tekstu, onda dobija ne-nultnu vrednost u tekst-vektoru pored dimenzije odgovarajućeg termina. Pošto bilo koji tekst sadrži ograničen broj termina (rečnik može imati milione termina) većina vektora teksta su dosta oskudni. Većina vektorskih sistema operiše u pozitivnom delu vektorskog prostora, to jest ni jednom terminu se ne dodeljuje negativna vrednost.



Slika 1 – Model vektorskog prostora

Da bi dodelio numeričku ocenu dokumentu za upit, model meri *sličnost* između vektora upita (i upit je običan tekst i može se pretvoriti u vektor) i vektora dokumenta. Definicija sličnosti između ova dva vektora ne postoji u modelu, ali se uglavnom koristi ugao između vektora kao mera razmaka između njih i kosinus ugla se koristi kao numerička sličnost (pošto kosinus ima korisnu sposobnost da je 1.0 za identične vrednosti i 0.0 za ortogonalne vrednosti). Kao alternativa, skalarni proizvod između dva vektora se koristi za merenje sličnosti. Ako su svi vektori forsirani da budu veličine jedinice, onda je kosinus ugla između dva vektora isti kao i njihov skalarni proizvod. Ako $\vec{D}$ predstavlja vektor dokumenta i $\vec{Q}$ predstavlja vektor upita, onda sličnost (ili ocena) dokumenta D za upit Q se može prikazati kao:

$$(\vec{D}, \vec{Q}) = \sum_{t_i \in Q, D} \omega_{t_{iQ}} \cdot \omega_{t_{iD}}$$

gde je $\omega_{t_{iQ}}$ vrednost od i -tog komponenta u upitnom vektoru $\vec{Q}$, i $\omega_{t_{iD}}$ je i -ti komponent u vektoru dokumenta $\vec{D}$. Pošto svaka reč koja nije prisutna u upitu ili dokumentu ima $\omega_{t_{iQ}}$ ili $\omega_{t_{iD}}$ vrednost 0, možemo da uradimo samo zbir termina koji se nalaze u upitu i dokumentu. Kako doći do $\omega_{t_{iQ}}$ i $\omega_{t_{iD}}$ nije definisano u modelu, ali je veoma kritično u efikasnosti pretraživanja sistema za pretraživanje informacija. $\omega_{t_{iD}}$ se često naziva *težina* termina i od dokumenta D .

2.2 Modeli verovatnoće

Modeli verovatnoće su bazirani na generalnom principu da dokumenti u kolekciji treba da rankiraju po opadajućoj verovatnoći njihove relevantnosti za upit.[7] Pošto prave verovatnoće nisu dostupne sistemu za pronalaženje informacija, modeli verovatnoća procenjuju verovatnoću relevantnosti dokumenta za određeni upit. Ova procena je ključni deo modela, i ovde je gde se većina modela verovatnoće razlikuju jedan od drugog. Početnu ideju o modelu verovatnoće su predložili Maron i Kuhns u radu objavljenom 1960.[8] Od tada, mnogi modeli verovatnoće su predloženi, svaki baziran na drugačijoj proceni verovatnoće.

Doduše, svi ovi modeli imaju osnovnu bazu. Označićemo verovatnoću relevantnosti nekog dokumenta D kao $P(R|D)$. Pošto je kriterijum rankiranja monoton pod transformacijom logaritamskog deljenja, možemo da rankiramo dokumenta kao $\log \frac{P(R|D)}{P(\hat{R}|D)}$, gde $P(\hat{R}|D)$ je verovatnoća da je dokument ne-relevantan. Jednostavnom Bajesovom transformacijom ovo možemo pretvoriti u $\log \frac{P(R|D) \cdot P(R)}{P(\hat{R}|D) \cdot P(\hat{R})}$

Predpostavljajući da predhodna verovatnoća relavance $P(R)$ je nezavisna od određenog dokumenata, i dakle konstantna u svim dokumentima, $P(R)$ i $P(\hat{R})$ samo skalarni faktori za konačne ocene dokumenata i mogu se izbaciti iz formule (u cilju rankiranja). Ovo dodatno pojednostavljuje formulu u $\log \frac{P(D|R)}{P(D|\hat{R})}$. Bazirani na pretpostavkama od procene $P(D|R)$, različiti modeli verovatnoće nastaju. U najjednostavnijoj verziji ovog modela, predpostavljamo da su termini (uglavnom reči) uzajamno nezavisni, i $P(D|R)$ se transformiše kao proizvod individualnih verovatnoća termina, to jest verovatnoću prisustva i odsustva termina u relevantnim i ne-relevantnim dokumentima:

$$P(D|R) = \prod_{t_i \in Q, D} P(t_i|R) \cdot \prod_{t_j \in Q, \hat{D}} (1 - P(t_j|R))$$

U ovoj formuli t_i označava verovatnoću prisustva termina u relevantnim dokumentima za sve termine koji se nalaze u upitu i dokumentu, a t_j verovatnoću odsustva termina iz relevantnih dokumenata za sve termine koje se nalaze u upitu i ne nalaze u dokumentu.

Ako p_i označava $P(t_i | R)$, i q_i označava $P(t_i | \hat{R})$, formula za rankiranje $\log \frac{P(D|R)}{P(D|\hat{R})}$ se redukuje na:

$$\log \frac{\prod_{t_i \in Q, D} p_i \cdot \prod_{t_j \in Q, \hat{D}} (1 - p_j)}{\prod_{t_i \in Q, D} q_i \cdot \prod_{t_j \in Q, \hat{D}} (1 - q_j)}$$

Za određeni upit, možemo dodati na ovo konstantu $\log \left(\prod_{t_i \in Q} \frac{1 - q_i}{1 - p_i} \right)$ da bi transformisali formulu da koristi samo termine koji se nalaze u dokumentu:

$$\log \prod_{t_i \in Q, D} \frac{p_i \cdot (1 - q_i)}{q_i \cdot (1 - p_i)}$$

ili

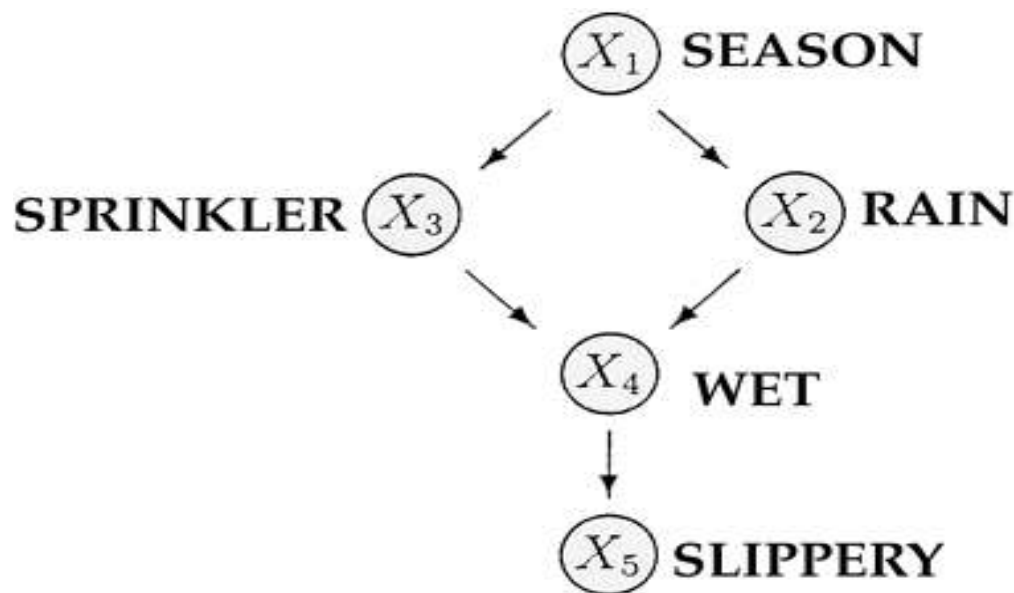
$$\sum_{t_i \in Q, D} \log \frac{p_i \cdot (1 - q_i)}{q_i \cdot (1 - p_i)}$$

Drugačije pretpostavke da procenu p_i i q_i daju drugačije funkcije za rankiranje dokumenata. Na primer, C. W. Croft i D. J. Harper [9] u svom radu pretpostavljaju da je p_i isto za sve termine u upitu i da je $\frac{p_i}{1 - p_i}$ konstanta koja se može ignorisati u cilju rankiranja. Oni takođe pretpostavljaju da skoro sva dokumenta u kolekciji nisu relevantna za određeni upit (što je blizu istini jer su kolekcije uglavnom veoma velike) i procenjuju q_i kao $\frac{n_i}{N}$ gde je N veličina kolekcije i n_i je broj dokumenata koji sadrže termin i . Ovim dobijamo funkciju za ocenjivanje $\sum_{t_i \in Q, D} \log \frac{N - n_i}{n_i}$.

Primećujemo da ako zamislamo $\log \frac{p_i \cdot (1 - q_i)}{q_i \cdot (1 - p_i)}$ kao težinu termina i u dokumentu D ovo formulisanje postaje veoma slično formulisanju sličnosti u modelu vektorskog prostora sa terminima upita kojima je dodeljena težina jedinice.

2.3 Model mreže zaključivanja

U ovom modelu, pronalaženje dokumenata je modelovano kao proces zaključivanja u mreži zaključivanja.[10] Većina tehnika koju sistemi za pronalaženje informacija koriste mogu da se implementiraju pod ovim modelom. U najjednostavnijoj implementaciji ovog modela, dokument instancira termin sa određenom jačinom, i kredit od više termina je skupljen od određenog upita da izračuna jednakost numeričke ocene za određeni dokument. Iz operacijske perspektive, jačina instanciranja termina od dokumenta može da se smatra kao *težina* termina u dokumentu, i rankiranje dokumenata u najjednostavnijoj formi ovog modela postaje slično rankiranju u modelu vektorskog prostora kao i modelima verovatnoće. Jačina instanciranja termina iz dokumenta nije definisana u modelu, i bilo koje formulisanje se može koristiti.



Slika 2 – Jednostavna Bajesova mreža zaključivanja. Opisuje veze između pet promenljivih (sezona u godini, da li pada kiša, da li je prskalice uključena, da li je trotoar mokar i klizav)

3 IMPLEMENTACIJA

Većina današnjih sistema za pronalaženje informacija su bazirani na strukturi podataka zvanoj *preokrenute liste*. To omogućava brzu dostupnost listi dokumenata koji sadrže termin sa ostalim informacijama (na primer, težinu termina u svakom dokumentu, relativne pozicije termina u svakom dokumentu itd.). Tipična preokrenuta lista se može predstaviti kao:

$$t_i \rightarrow \langle d_a, \dots \rangle, \langle d_b, \dots \rangle, \dots, \langle d_n, \dots \rangle,$$

što predstavlja da termin i se sadrži u $d_a, d_b, \dots, d_n$ i skladišti bilo koju drugu informaciju. Svi modeli mogu da se implementiraju koristeći preokrenute liste. One koriste činjenicu da za određeni upit većina sistema za pronalaženje informacija ocenjuju samo mali broj dokumenata koji sadrže neki termin iz upita. To omogućava sistemu da ocenjuje samo dokumente koji imaju ne-nultnu numeričku ocenu. Većina sistema drži ocene dokumenata u strukturi podataka zvanoj *gomila* i na kranu procesiranja vraćaju najviše ocene dokumenata za određeni upit. Pošto su svi dokumenti označeni po terminima koje sadrže, proces generisanja, pravljenja i čuvanja reprezentacija dokumenata se zove *indeksiranje* i preokrenuti fajlovi rezultata se zovu *preokrenut indeks*.

Većina sistema za pronalaženje informacija koriste jednu reč kao termin. Reči za koje se smatra da nisu informativne, kao na primer funkcijske reči (u, iz, od ...) se često ignorišu. Spajanje raznih formi iste reči sa njenim korenom, (što se zove *kalemljenje* u žargonu) takođe koriste mnogi sistemi. Glavna ideja kalemljenja je da korisnici koji traže informaciju o *pretraživanju* će isto biti zainteresovani u dokumenta koja imaju informaciju o *pretražiti, pretražio, pretraživati, pretraživač*. Ovo takođe pravi sistem dosta podložan greškama zbog lošeg kalemljenja. Na primer, korisnik zainteresovan za *put autobusa* može da dobije dokument nazvan *putnik autobusa* zbog kalemljenja. Dosta ovakvih programa je napravljeno za razne jezike tokom godina, svaki sa svojom pravilima. Doduše, korisnost kalemljenja za poboljšano pretraživanje se uvek razmatralo, pogotovo za engleski jezik. U njemu, po proseku kalemljenje daje mala poboljšanja u efikasnosti pretraživanja, dok u slučajevima gde uzrokuje lošu pretragu korisnik će biti dosta iznerviran.[11] Kalemljenje je uglavnom korisnije za jezike koji imaju mnogo modulacija reči (kao nemački na primer).

Neki sistemi koriste fraze od više reči kao termine indeksa. Pošto se fraze smatraju značajnijim od jedinstvenih reči, pronalazak fraze u dokumentu se smatra informativnijim nego pronalaženje jedne reči. Izučavano je nekoliko tehnika za generisanje lista fraza, od kompletno lingvističkih (baziranih na raščlanjivanju rečenica) do kompletno statističkih (baziranih na brojanju ponavljanja reči). U istraživačkoj zajednici je prihvaćeno da su fraze važne jedinice indeksiranja i da poboljšavaju efektivnost pretraživanja, mada stil generisanja fraza nije mnogo bitan. Studije koje su upoređivale performanse lingvističkih fraza sa statističkim frazama nisu našle nikakvu razliku.[12]

3.1 BM25 rankirajuća funkcija

Jedna od praktičnih primera implementacije je BM25 rankirajuća funkcija koja je produkt Okapi projekta (Stephen Robertson, Karen Sparck Jones)[13], koja je korišćena u pretraživačima kao što je Lucene projekat.

Rankirajuća funkcija koristi određeni dokument i upit i kao rezultat vraća numeričku ocenu, pri čemu najrelevantniji dokumenti imaju najveću ocenu. U BM25 funkciji rezultat je linearna kombinacija težina ocena za svaku reč u upitu. Tri faktora određuju težinu upitnog termina:

- Frekvencija termina – Broj pojavljivanja reči iz upita u određenom dokumentu (term frequency - TF)
- Obrnuta frekvencija dokumenta od termina – Reči koje se nalaze u mnogo dokumenata imaju veliku frekvenciju dokumenta, to jest malu obrnutu frekvenciju dokumenta i smatra se da imaju malu relevantnost za upit (inverse document frequency - IDF)
- Dužina dokumenta – Ogromni dokumenti mogu da sadrže sve reči iz upita a da nemaju veze sa upitom, zato su kraći dokumenti koji sadrže reči iz upita mnogo relevantniji

BM25 funkcija uzima sve ove elemente u obzir. Predpostavimo da smo kreirali indeks od N dokumenata da bi izračunali $TF(q_i, d_j)$, broj koliko puta se reč q_i sadrži u dokumentu d_j . Takođe predpostavimo tabelu za prebrojavanje frekvencije dokumenta, $IDF(q_i)$, koja daje broj dokumenata koji sadrže reč q_i . Onda za određeni dokument d_j i za upit koji sadrži reči $q_1:N$, imamo :

$$BM25(d_j, q_1:N) = \sum_{i=1}^N IDF(q_i) \cdot \frac{TF(q_i, d_j) \cdot (k + 1)}{TF(q_i, d_j) + k \cdot \left(1 - b + b \cdot \frac{|d_j|}{L}\right)}$$

gde je $|d_j|$ dužina dokumenta d_j u rečima, a L je prosečna dužina dokumenata u celom skupu dokumenata $L = \sum_i \frac{|d_i|}{N}$. Imamo dva parametra k i b , koje možemo podesiti ukrštenom validacijom, tipične vrednosti su $k = 2.0$ i $b = 0.75$. $IDF(q_i)$ je obrnuta frekvencija dokumenta od reči q_i koju dobijamo od :

$$IDF(q_i) = \log \frac{N - DF(q_i) + 0.5}{DF(q_i) + 0.5}$$

Naravno ne bi bilo praktično da primenimo BM25 funkciju za svaki dokument u skupu dokumenata. Zato sistemi za pronalaženje informacija kreiraju indeks koji nabroja, za sve reči u rečniku, dokumente koji sadrže tu reč. Ovo nazivamo *listom pogodaka* za određenu reč. Onda kada dobijemo upit, ukrstimo liste pogodaka za reči u upitu i samo ocenjujemo ta dokumenta.

3.2 PageRank algoritam

PageRank algoritam je jedna od dve originalne ideje koje su odvojile Google od drugih web pretraživača (druga ideja je korišćenje usidrenjog teksta -podvučenih reči u hiperlinku- za indeksiranje stranica, iako usidrenjeni tekst nije bio na stranici koja se indeksirala).[19] PageRank je izmišljen da bi rešio probleme *TF* ocena. Na primer, ako imamo upit [Nike], kako da budemo sigurni da je www.nike.com (glavna web stranica poznatog proizvođača patika) prvi rezultat iako neka druga stranica pominje termin “Nike” više puta? Ideja je da www.nike.com ima mnogo veza (link) na stranicu, pa bi trebao biti rangiran više: svaka veza je ocena kvaliteta povezane stranice. Ali ako bismo računali samo veze na stranicu onda bi bilo moguće da neko kreira mnogo praznih stranica koje će se vezati za određenu stranicu i time povišavati njenu ocenu. PageRank algoritam je dizajniran da ocenjuje veze iz sajtova velikog kvaliteta više nego veze iz standardnih sajtova. Sajt velikog kvaliteta definišemo kao sajt koji je povezan od drugih sajtova velikog kvaliteta. PageRank za stranicu p definišemo kao :

$$PR(p) = \frac{1 - d}{N} + d \sum_i \frac{PR(in_i)}{C(in_i)}$$

gde je $PR(p)$ PageRank od stranice p , N je broj stranica u skupu dokumenata, in_i su stranice koje se vezuju na stranicu p , a $C(in_i)$ je broj odlazećih veza na stranici in_i . Konstantu d možemo objasniti kroz *model nasumičnog web surfera*: zamislimo web surfera koji kreće da surfuje od neke nasumične strane. Sa verovatnoćom d surfer klikće na neku vezu na stranici (jednako ih bira), i sa verovatnoćom $1-d$ mu dosadi stranica i ponovo kreće surfovanje sa nasumične stranice na web-u. PageRank od stranice p je onda verovatnoća da će nasumični web surfer biti na toj stranici u bilo kom trenutku. PageRank se može izračunati iterativnom procedurom: počecemo sa $PR(p)=1$, i ponavljati algoritam sve dok rezultat ne konvergira.

3.3 HITS algoritam

Hyperlink-Induced Topic Search algoritam ili HITS je još jedan algoritam baziran na analizi veza između sajtova.[19] HITS se razlikuje od PageRank algoritma na više načina. Za početak, rankira stranice u zavisnosti od upita. To znači da se mora ponovo računati za svaki upit, što zahteva dosta računarske moći i zato većina pretraživača izbegava ovaj pristup. Za određeni upit, HITS prvo pronalazi set stranica koje su relevantne za upit. To radi tako što ukršta liste pogodaka od reči u upitu, i onda dodaje stranice u komšiluk veza tih stranica (stranice koje se vežu ili su vezane od neke stranice iz originalnog seta relevantnih stranica).

Svaka stranica u ovom setu se smatra *autoritetom* za upit, ako ostale stranice u relevantnom setu pokazuju na nju. Određena stranica se smatra *čvorom* ako pokazuje na druge autoritativne stranice u relevantnom setu. Kao i kod PageRank-a ne želimo da samo računamo broj veza, želimo da damo veću vrednost čvorovima i autoritetima velikog kvaliteta. Napravićemo proces koji će unaprediti proces ocenjivanja autoriteta stranice da bude suma ocena čvorova od stranica koje pokazuju na nju, i ocena čvora da bude suma ocena autoriteta od stranica na koje pokazuje. Ako onda normalizujemo ocene i ponovimo k puta, proces će se konvergirati.

funkcija HITS(*upit*) **vraća** stranice sa bojevima čvorova i autoriteta

```
stranice ← PROŠIRI_STRANICE(RELEVANTNE_STRANICE(upit))
za svako  $p$  u stranice uradi
     $p.AUTORITET ← 1$ 
     $p.ČVOR ← 1$ 
ponovi dok konvergiranje uradi
    za svako  $p$  u stranice uradi
         $p.AUTORITET ← \sum_i DO - VEZA_i(p).ČVOR$ 
         $p.ČVOR ← 1 \sum_i OD - VEZA_i(p).AUTORITET$ 
    NORMALIZUJ(stranice)
vraća stranice
```

Primer HITS algoritma za idračunavanje čvorova i autoriteta za određeni upit.

RELEVANTNE_STRANICE vraća stranice koje odogovaraju upitu, i PROŠIRI_STRANICE dodaje svaku stranicu koja se veže ili je vezana od neke relevantne stranice. NORMALIZUJ deli ocenu svake stranice sa sumom korena ocena svih stranica (odvojeno za čvorove i autoritete)

4 OCENJIVANJE SISTEMA

Objektivno ocenjivanje efektivnosti pretrage sistema za pretraživanje informacija je temelj ovakvih sistema. Napredak u ovoj naučnoj oblasti kritično zavisi od eksperimentisanja sa novim idejama i ocenjivanju efekta tih ideja. Još u prvim godinama istraživanja, istraživačima u zajednici je bilo jasno da će objektivno ocenjivanje tehnika pretrage imati ključnu ulogu u polju. Testovi koji su se vršili u Cranfield Univerzitetu, 1960-ih godina,[4] su uspostavili set željenih karakteristika za sistem za pretraživanje informacija. Iako je bilo nekoliko rasprava tokom godina, dve željene osobine su prihvaćene kao mera efektivnosti pretrage : *opoziv* i *preciznost*.

Objasnićemo ove pojmove preko sledeće tabele:

	U setu rezultata	Nisu u setu rezultata
Relevantni	30	20
Nerelevantni	10	40

Tabela 1 – primer pretrage

Zamislimo da sistem za pronalaženje informacija vrati set rezultata za određeni upit, za koji znamo koji dokumenti su relevantni a koji nisu, iz skupa od 100 dokumenata. Preciznost meri proporciju dokumenata u setu rezultata koji su relevantni. U našem primeru preciznost je $30/(30+10)=0.75$. Mera lažne pozitivnosti je $1-0.75=0.25$. Opoziv meri proporciju svih relevantnih dokumenata u kolekciji koji su u setu rezultata. U našem primeru opoziv je $30/(30+20)=0.6$. Mera lažne negativnosti je $1-0.6=0.4$.

Dobar sistem za pronalaženje informacija bi trebao da pronađe što više relevantnih dokumenata (tj. da ima veliki opoziv), i što manje nerelevantnih dokumenata (tj. veliku preciznost). Nažalost ova dva cilja su uglavnom kontradiktorna. Tehnike koje unapređuju opoziv slabe preciznost i obrnuto. I opoziv i preciznost su set-orijentisane mere i nemaju pojam rankiranog pretraživanja. Istraživači su koristili nekoliko varijanti opoziva i preciznosti da ocene rankirano pretraživanje. Na primer ako dizajneri sistema misle da je preciznost više važna njihovim korisnicima, mogu da mere preciznost na 50% opoziva, što bi pokazalo koliko nerelevantnih dokumenata bi korisnik morao da pročita da bi našao pola relevantnih. *Prosečna preciznost* je mera sa jednom vrednošću koja je najčešće koristi za ocenjivanje rankiranog pretraživanja. Prosečna preciznost se meri tako što se izmeri preciznost u nekoliko tački opoziva (na primer 10% , 20%, 30% itd) i onda se izračuna prosek.[14]

5 MODIFIKACIJA UPITA

U ranim godinama istraživanja, istraživači su shvatili da je teško za korisnike da naprave efektivne upite. Mislilo se da dodavanjem sinonima od upitnih reči u upit se može unaprediti efektivnost pretrage. Početno istraživanje se oslanjalo na knjigu sinonima za nalaženje sinonima,[15] iako je bilo dosta teško naći dobru generalnu knjigu sinonima. Istraživači su razvili tehnike za automatsko generisanje knjiga sinonima za modifikovanje upita. Većina automatskih metoda su bazirani na analiziranju učestalosti više istih reči u dokumentu. Većina tehnika modifikacija upita baziranih na automatskom generisanju knjiga sinonima nisu imale uspeha u unapređivanju efektivnosti pretrage. Glavni razlog ovoga je nedostatak konteksta upita u procesu modifikacije. Nisu sve reči koje se odnose na upit značajne za kontekst upita.

U 1965 Rocchio je predložio korišćenje povratnih informacija o relevantnosti za modifikaciju upita.[16] Ideja povratih informacija o relevantnosti je da je lako korisnicima da osude da li su dokumenta koja su dobili iz pretrage relevantna ili ne za njihov upit. Koristeći ove procene relevantnosti, sistem može automatski da generiše bolji upit (na primer dodavajući nove srodne termine) za dalju pretragu. Generalno, korisnik se pita da oceni relevantnost samo od par dokumenata na vrhu pretrage. Bazirano na ovim ocenama, sistem modifikuje upit i pravi novi upit da bi pronašao relevantnija dokumenta iz kolekcije. Pri testiranju, povratne informacije o relativnosti su se pokazale kao veoma efektivne za poboljšavanje pretrage.

Nove tehnike za modifikaciju upita bez ikakve pomoći korisnika su se razvijale u ranim 1990-ima. Najvažnija tehnika je korišćenje pseudo povratnih informacija o relativnosti, varijanta od povratnih informacija o relativnosti.[17] Pošto su par najviših dokumenata dobijenih od sistema za pronalaženje informacija uglavnom o generalnoj temi upita, uzimanje srodnih termina iz ovih dokumenata bi trebalo da donese korisne nove termine bez obzira na relevantnost dokumenta. U ovoj tehnici, sistem predpostavlja da su par najviših dokumenata dobijenih iz određenih upita relevantni za taj upit, i pravi informacije o relativnosti za generisanje novog upita. Ovaj proširen novi upit se koristi za rankiranje dokumenata koji će se pokazati korisniku. Pseudo povratne informacije su se pokazale kao veoma efektivne tehnika, posebno za kratke upite.

5.1 Upiti u obliku pitanja

Pronalaženje informacija je zadatak pronalaženja dokumenata koji su relevantni za upit, gde upit može biti pitanje, ili samo oblast ili koncept teme. Odgovaranje na pitanja je malo drugačiji zadatak, u kome je upit pitanje, a odgovor nije rankirana lista dokumenata nego kratak odgovor – rečenica ili samo fraza. Postojalo je dosta sistema za odovaranje na pitanja baziranih na obrađivanju prirodnog jezika još od 1960-tih godina, ali tek su od 2001 godine ti sistemi počeli da koriste Web pronalaženje informacija da značajno povećaju širinu pretrage.

AskMSR sistem [18] je tipičan Web baziran sistem za odgovaranje na pitanja. Baziran je na ideji da će većina pitanja biti odgovorena mnogo puta na Web-u, tako da odgovaranje na pitanja se smatra problemom preciznosti a ne opoziva. Ovakav sistem ne mora da brine o drugačijim načinima na koje se može postaviti neko pitanje, samo treba da nađe jedno od njih. Zamislamo da za upit [Ko je ubio Abrahama Linkolna?] sistem ima samo pristup jednoj enciklopediji gde piše “Dzon But je promenio istoriju sa metkom. On će biti upamćen kao čovek koji je završio život Abrahama Linkolna”. Da bi iskoristili ovaj pasus da odgovorimo na pitanje, sistem bi morao da zna da završiti život može da znači ubistvo, da “On” se odnosi na But-a i druge lingvističke i semantičke činjenice.

AskMSR ne zna ništa o zamenicama, ili o ubijanju ili bilo kom drugom glagolu, ali zna razne vrste pitanja i kako se mogu modifikovati da bi bili upiti za pretraživač. Zna da pitanje [Ko je ubio Abrahama Linkolna?] može da se izmeni u upit [* je ubio Abraham Linkoln] i [Abraham Linkoln je ubijen od *]. Sistem pokreće ove izmenjene upite i pregledava rezultate, ne cele Web stranice nego samo kratak tekst koji se pojavljuje blizu upitnih termina. Ovi rezultati se dele u 1-, 2- i 3- grame i evidentiraju se za frekvenciju u setu rezultata i za težinu: n -gram koji dolazi iz specifičnog izmenjenog upita (na primer tačna fraza odgovora [Abraham Linkoln je ubijen od *]) bi dobio veću težinu nego od generalno izmenjenog upita, kao [Abraham ILI Linkoln ILI ubijen].

Kad se n -grami ocene, filtriraju se po očekivanom tipu. Ako originalni upit počinje sa “ko” onda filtriramo po imenima ljudi, za “koliko” filtriramo po brojevima, za “kada” po datumu ili vremenu. Takođe postoji filter koji kaže da odgovor ne bi trebao da bude deo pitanja; zajedno ovi mehanizmi bi trebalo da daju “Dzon But” (a ne “Abraham Linkoln”) kao najavše ocenjeni rezultat.

6 PRETRAŽIVANJE SLIKA BAZIRANO NA TEKSTU

Pretraživači slika bazirani na tekstu se fokusiraju na pretraživanje uz pomoć teksta povezanog sa slikom ili u vezi sa tom slikom. Za ovakve pretraživače uglavnom se koriste *metapodaci* za opis samih slika. Metapodaci se mogu definisati kao podaci o podacima. Metapodaci slika mogu biti raznog tipa, to jeste tagovi, ključne reči i opisi veza određene slike. U to spadaju podaci dodati od same kamere (vreme, datum i GPS koordinate), reči dodate od samih korisnika da bi objasnili sliku (tag) ili automatske anotacije koje dodaje sistem za pretraživanje radi poboljšanja pretrage i indeksiranja (mogu se zvati i *auto-anotacije* ili lingvističko indeksiranje) [21]. Neki sistemi za pretraživanje bazirani na tekstu kombinuju metapodatke drugačijeg konteksta iz raznih izvora, radi dopunjivanja metapodataka određene slike [22]. Takodje radi dodavanja semantike tagovima, tagovi se se mapiraju u *ontologije*. Ontologija je viši nivo opisa od termina i sadrži semantičko značenje određene reči.

Pretraživači slika bazirani na tekstu bolje odgovaraju današnjim potrebama nego pretraživači slika bazirani na sadržaju zbog sve većeg broja slika u kolekcijama. Ovakvi sistemi troše manje kompjuterskih resursa i mnogo su kompatibilniji sa ljudskim konceptima što se tiče upita nego sistemi za pretraživanje slika baziranim na sadržaju.

Kontekst je bilo koja informacija koja se može iskoristiti da karakteriše situaciju određenog entiteta. U većini sistema za pretraživanje slika entitet je uglavnom uređaj koji pravi slike. Tipovi konteksta koji mogu pomoći u karakterizaciji situacije na slici su GPS lokacija i datum i vreme. Dodavajući metapodatke na samu sliku, može se povećati znanje o slici i takođe se može smanjiti semantički raskorak veoma efikasno. Informacija o lokaciji ponekad sama može dati semantički sadržaj slici.

Ručno dodati tagovi mogu biti veoma korisni za sistem da pretraživanje slika, ako su dostupni. To su ključne reči dodate slici od strane individualnih korisnika. U teoriji oni reprezentiraju prirodnu percepciju slike od individualnog korisnika. Jedan problem sa ručno dodatim tagovima je to što su ljudski tagovi *subjektivni*, što znači da individualni korisnici imaju drugačije percepcije, i koriste drugačije tagove da opišu slike. Parametri za ovakvo tagovanje mogu obuhvatiti faktore kao što su socijalni, kulturni, kontekst, vreme itd. Takođe *tagovanje sa saradnjom korisnika* je veoma popularno danas na sajtovima za prosleđivanje slika i društvenim mrežama, što pomaže istraživačima da razumeju ljudsku percepciju i da prilagode šablone tagovanja za druge slike.

Pošto moderna tehnologija pravi uređaje za pravljenje slika (kamere i mobilne telefone) sve više dostupnim, mobilnim i pristupačnim, korišćenje uređaja za pravljenje slika se eksponencijalno povećava. Takođe ljudi dele više slika između sebe preko sajtova i društvenih mreža nego ikada ranije. Iako je pretraživanje slika bazirano na tekstu prikladno za velike količine slika koje su dostupne, zadatak pretraživanja njih, tagova, individualnih oblika može istrošiti dosta vremena i resursa. Zbog korisnički specifičnih i nedoslednih metapodataka za slike u velikim kolekcijama slika, upiti mogu dovesti do rezultata ogromne veličine. Dakle, važno je smanjiti obim pretrage, fokusirajući se samo na najrelevantnije slike. Tehnike pretraživanja baziranim na tekstu su dobro utvrđene u oblasti *pronalaženja informacija*. Ove tehnike su usavršavane decenijama sa ciljem filtriranja najvažnijih tekstualnih dokumenata u relaciji sa specifičnim upitima od strane korisnika. Pretraživanje slika bazirano na tekstu ima veliki potencijal za prilagođavanje tehnika koje se koriste u sistemima za pronalaženje informacija.

6.1 Selekcija obima pretrage

U teoriji informacije, tehnike za smanjivanje obima pretrage su dobro utvrđene. Analiza glavnog komponenta (AGK), smanjuje obim pretrage tako što analizira sekvencu (na primer tekst dokument) , i otkriva unutrašnju strukturu podataka u njoj i opisuje informacije iz najboljeg pogleda. AGK je jedna instanca klase od analitičkih algoritama zvanih analize zajedničkih faktora, koji traže najmanji broj dimenzija [23].

Odlike ili svojstva slike mogu se odrediti važnim iz više razloga, na primer specifičnost područja, specifičnost diskriminacije (na primer boja kose u pretraživanju lica), relevantnost korisnika, relevantnost lokacije itd [20]. Najjednostavniji način smanjivanja obima pretrage je pomoću *frekvence termina*. U teoriji informacije, ovaj metod se uglavnom koristi za prebrojavanje svake reči u dokumentu, da bi se odredilo koje reči se najviše koriste i ocenili dokumenti po ključnoj reči, u zavisnosti od frekvence ključne reči u dokumentu [20].

Ocenjivanje metapodataka na ovaj način je veoma bitno za pretraživanje slika bazirano na tekstu i može se koristiti za lociranje termina sa velikom reprezentacijom. Jedna mana metoda frekvence termina je to što ne dobija semantički smisao dokumenta kao AGK.

6.2 Konstrukcija indeksa

Da bi dokumenti (u ovom slučaju slike) mogli da se pretražuju moraju se prvo indeksirati. Dobijena *sekvenca karaktera* se *tokeniše*, to jest deli se na više tokena sa time što se neki karakteri kao znakovi interpunkcije odbacuju. *Tokeni* su termini ili reči u dokumentu. Na primer:

Ulaz: Šetnja na plaži sa društvom.

Izlaz: šetnja na plaži sa društvom

Nakon tokenisanja odbacuju se uobičajeni termini. Ovi termini veoma malo pomažu korisniku da dobije ono što traži i uglavnom se nazivaju *stop rečima*. Oni se mogu naći ili preko filtriranja najučestalijih termina preko frekvence termina ili ručno napraviti listu učestalih reči koje će se odbaciti. Primer nekih stop reči koje treba odbaciti: jedan, i, jesu, kao, od, za, u, itd.

U mnogo situacija tokeni (reči) nisu indentični, ali je ipak poželjno da se smatraju parom. Na primer ako se pretražuje termin Wi-Fi, poželjno je naći dokumente koje sadrže Wi-Fi i Wi-Fi, ili ako se pretražuje termin MUP poželjno je naći dokumente koje sadrže MUP i M.U.P. *Normalizacija tokena* je proces pretvaranja tokena da bi se ovakvo pronalaženje desilo, bez obzira na razlike u sekvenci karaktera. Takođe je jako važno za normalizaciju transformisati sve karaktere u mala slova da bi pretraga termina Srbija i srbija bila jednaka.

Dokumenti će sadržati drugačije forme od iste reči, na primer organizovati, organizuje, organizovano. Isto tako, postoje gramatički drugačije reči sa sličnim značenjem (demokratija, demokratski, demokratizacija). Cilj *lematizacije* je da redukuje ovakve forme reči radi lakšeg pretraživanja:

čovek, ljudi, dečak, dečaci → muškarac

6.3 Frekvencija termina – Frekvencija obrnutih dokumenata (FT-FID)

Kada je dokument spreman za indeksiranje, on se uglavnom ocenjuje u skladu sa njegovom relevantnošću za razne upite. Frekvencija Termina – Frekvencija obrnutih dokumenata je jedan takav metod. On se inače koristi u pretraživanju informacija i *kopanju teksta* [20]. *Frekvencija termina* (FT) se koristi pri ocenjivanju dokumenta u odnosu koliko se specifična reč ponavlja u relaciji sa totalnim brojem reči u dokumentu. Na primer ako dokument sadrži 100 reči i reč auto se ponavlja 5 puta u tekstu, frekvencija termina za auto u ovom dokumentu je:

$$FT(\text{auto}) = 5/100 = 0.05$$

Frekvencija obrnutih dokumenata (FOD) se koristi za ocenjivanje reči u odnosu sa svim dostupnim dokumentima. Ako malo dostupnih dokumenata sadrže specifičnu reč, FOD je veća nego ako puno dokumenata sadrže reč. Frekvencija obrnutih dokumenata se izračunava sa logaritamskom jednačinom koja uzima za obzir broj dostupnih dokumenata i koliko njih sadrži specifičnu reč. Na primer ako je dostupno 100 dokumenata i 5 njih sadrže reč auto onda je FOD termina auto:

$$FOD(\text{auto}) = \log(100/5) = 2.996$$

FT-FOD je proizvod *frekvencije termina* i *frekvencije obrnutih dokumenata* :

$$FT-FOD = FT * FOD$$

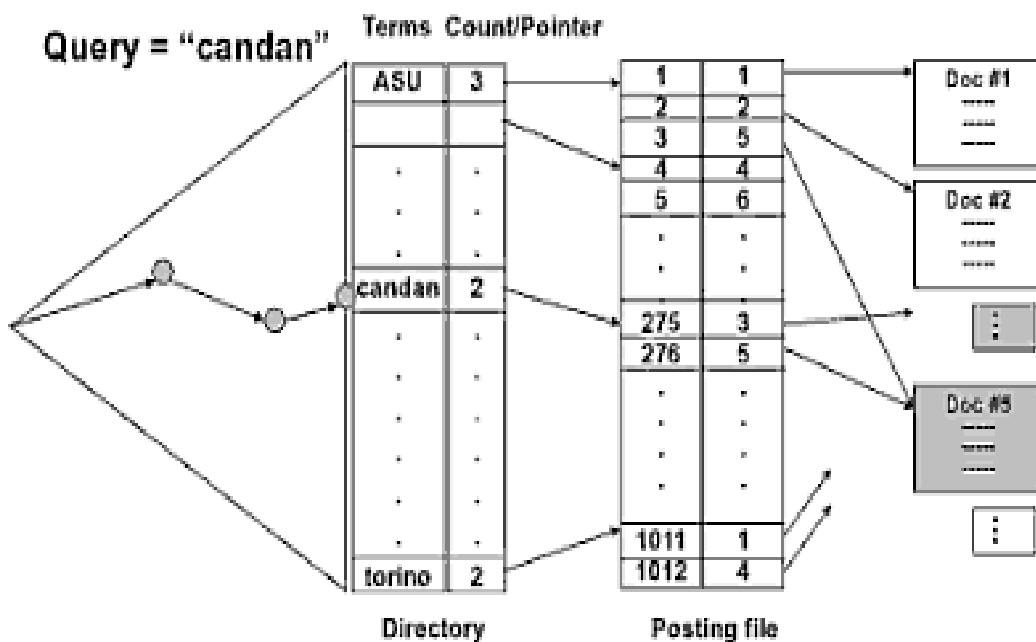
U slučaju termina auto za date dokumente, *frekvencija termina – frekvencija obrnutih dokumenata* je :

$$FT-FOD = (1) * (2) = 0.05 * 2.996 = 0.15$$

6.4 Indeksiranje obrnutih dokumenata

Indeks obrnutih dokumenata je struktura pretraživanja koja se uglavnom koristi u pronalaženju informacija, ali se takođe koristi i u obimnim sistemima za pretraživanje slika velikog obima zbog jako dobre efikasnosti. Ona sadrži mapiranje između termina u bazi podataka i lokaciju dokumenata koje sadrže te specifične termine. Struktura obrnutih dokumenata je sastavljena na sledeći način: [20]

- Direktorijum termina sadrži sve reči koje postoje u bazi podataka. Svaka reč sadrži relaciju na odgovarajuću obrnutu listu. Direktorijum takođe ima i broj dokumenata koji sadrže određenu reč.
- Obrnute liste se uglavnom čuvaju u javnom fajlu koji sadrži same relacije na dokumente.
- Struktura pretraživanja se uglavnom koristi za efikasno pristupanje direktorijumom obrnutih dokumenata, i njihovom pretraživanju po terminu upita. Strukture pretraživanja koje se široko koriste su hesh fajlovi za tačno uparivanje, B+ drveće, suffix automata. [20]



Slika 3 – Struktura obrnutih dokumenata, primer pretrage [20]

7 PRETRAŽIVANJE SLIKA BAZIRANO NA SADRŽAJU

Pretraživači slika bazirani na sadržaju se fokusiraju na analizi sadržaja slike. Sadržaj slike se može definisati kao njene boje, oblici, texture i bilo koja informacija koja se može izvesti iz same slike. Pošto slike generalno imaju ogroman sadržaj, a ljudska percepcija je jako subjektivna, ni jedan tekstualni opis slike ne može biti apsolutno tačan ili kompletan. Zbog toga su pretraživači slika bazirani na sadržaju efikasniji od pretraživača baziranih na tekstu.

Da bi sistemi za pretraživanje slika bili još efikasniji, oni ponekad koriste pomoć korisnika pri biranju relevantnih dokumenata iz početnog seta. Ovaj metod se zove *povratna informacija o relevantnosti*. Ideja povratne informacije o relevantnosti je da uzme rezultate od određenog upita i da iskoristi informaciju da li su ti rezultati relevantni za novi upit. Problem pretraživanja slika koristeći povratne informacije o relevantnosti je privukao jako veliku pažnju od akademskih istraživača. Sistem za pretraživanje slika koji bi se koristio u realnom svetu mora biti precizan, zahtevati minimalnu interakciju sa korisnikom, biti efikasan, sposoban za skaliranje na ogromne kolekcije slika i podržavati sesije sa više korisnika. Za dobru preciznost, potrebni su efektivni metodi za učenje relevance odlika slika baziranih od pomoći korisnika. Efikasnost i skaliranje zahtevaju dobru strukturu indeksiranja za dobijanje rezultata. Struktura indeksiranja mora da dozvoli da se relevantan odlik slika stalno menja u slopu sa novim upitima i povratim informacijama od strane korisnika [24].

Tačno rešenje ovog problema može voditi ka rešenju u mnogo različitih oblasti istraživanja. Neke od današnjih primena su bezbednost; poređenje slika za privatne firme za bezbednost kao i u istragama policije, arhivi medija; dobijanje i upravljanje vizuelnim asset-ima, naučne slike; dobijanje i klasifikacija slika u odnosu na specifičan vizuelni deo, umetničke kolekcije, medicinske dijagnoze, arhitektonski i inženjerski dizajn itd. U pretraživanju slika baziranom na sadržaju postoje uglavnom dva glavna pristupa: diskretan pristup i neprekidan pristup [25]. Diskretan pristup je inspirisan od strane pretraživanja tekstualnih informacija i koristi tehnike kao što su obrnuti fajlovi i metrički sistem za tekstualno pretraživanje. Ovaj pristup zahteva da se sve osobine mapiraju u binarne osobine; prisustvo određene osobine slike se tretira kao prisustvo reči u tekstualnom dokumentu. Neprekidan pristup je sličan klasifikaciji najbližeg komšije. Svaka slika je predstavljena od strane vektora osobina i te osobine se porede koristeći razna merenja razdaljine. Slike sa najmanjim razdaljinama se ocenjuju sa najvišom ocenom u procesu pretrage.

Procedure za pretraživanje slika se mogu podeliti u dve odvojene opcije: upit po tekstu i upit po primeru. U upitu po tekstu, upiti su tekst a objekat pretrage su slike. U upitu po primeru upiti su slike i objekat pretrage su slike. pristupi bazirani na tekstu primenjuju regularne tehnike za pretraživanje teksta na anotacije ili opise slike. pristupi bazirani na sadržaju primenjuju tehnike obrađivanja slika radi izvlačenja odlika slika i pronalaženje relevantnih slika. Cilj zadatka fotografske pretrage je da nađe što više relevantnih slika iz kolekcije za zadatu izjavu koja opisuje informaciju koju korisnik traži. Kada pretraživanje bazirano na tekstu i pretraživanje bazirano na slici sarađuju pri pretrazi to se zove fuzija ili višemodalno pretraživanje. Fuzija između pretraživanja baziranog na tekstu i pretraživanja baziranog na slici se može desiti u mnogo različitih momenata, i u zavisnosti od toga možemo imati veoma različita rešenja.

7.1 Predstavljanje slika i teksta

U pretraživanju slika postoje mnogo različitih tipova odlika koje se mogu izvući iz slika. Odlke se mogu kategorisati u sledeće tipove: predstavljanje boja, predstavljanje tekstura, lokalne odlike i predstavljanje oblika.

7.1.1 Histogrami boja

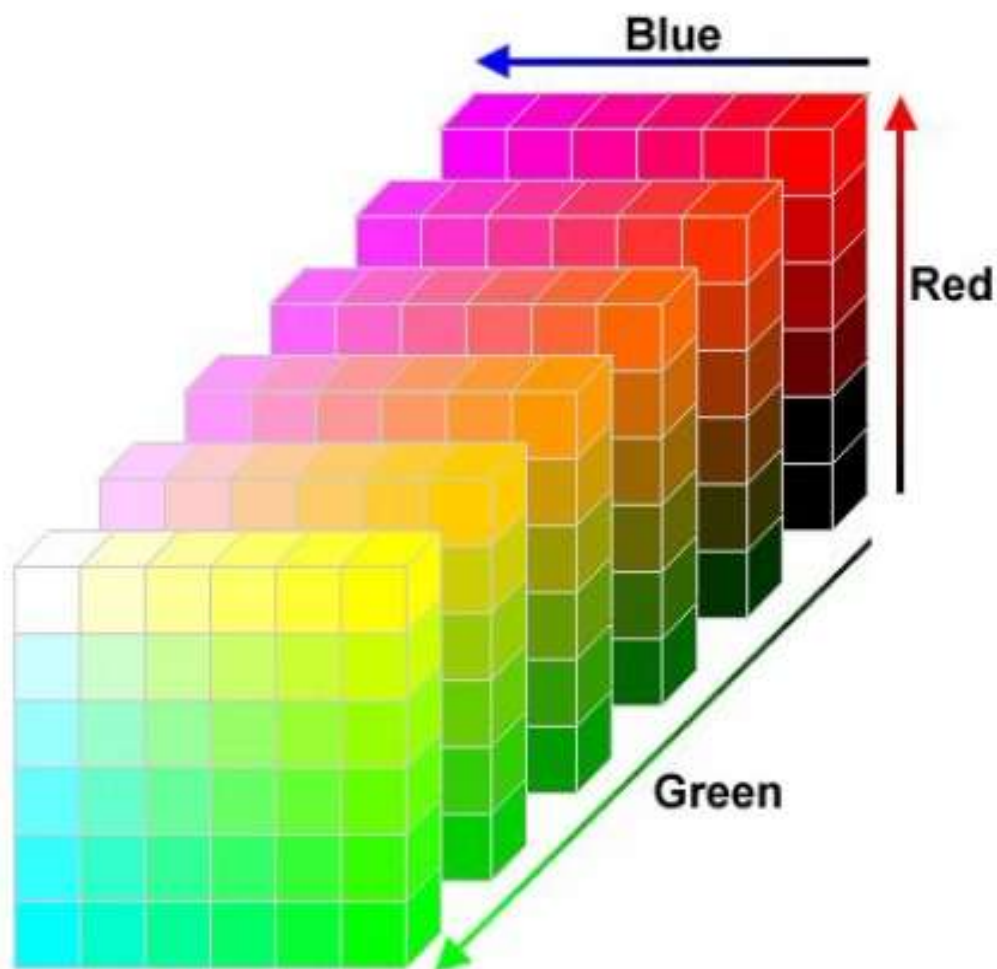
Jedan od najuobičajenih načina za predstavljanje slike je predstavljanje putem histograma boja. Svaka slika se može predstaviti kao vektor gde je svaka pozicija kvantitet boje u slici. Histogrami boja su jedni od najbazičnijih pristupa i osnov mnogih sistema za pretraživanje slika. Uprkos njihovoj jednostavnosti doprinose velikom preciznošću pretrage. Histogrami su nezavisni od veličine slika jer se svaki histogram normalizuje predhodno.

Neprebrojeni histogrami su tip histograma boja koji se najlakše dobijaju ali nisu veoma precizni. Kada je slika napravljena u matrici, konstruiše se vektor gde prvih 256 pozicija predstavljaju crvenu boju, sledećih 256 zelenu boju i poslednjih 256 plavu boju. Znači vektor će imati 768 pozicija. Na primer, ljubičasta boja sadrži crvenu zelenu i plavu boju tako da je prvi korak otkriti vrednost svake ove boje. Sledeći korak će biti da se ispita matrica boja i da se dodeli pozicija u konačnom vektoru za svaki piksel. Za ljubičasti piksel u primeru crvena, zelena i plava komponenta je izabrana i njena vrednost u vektoru je povećana. Ako se sve vrednosti u vektoru dodaju biće ih 3 puta više od ukupnog broja piksela u slici. Zadnji korak će biti da se normalizuje vektor deljenjem sa ukupnim brojem piksela, da bi vektor bio nezavisan od različitih veličina slika i da bih imao meru proporcije.

Ovo je jednako slaganju 3 različita histograma boje jedan za drugim. Glavni problem ovakvih histograma boje je to što se svaka komponenta boje tretira različito i to može dovesti do pogrešnih rezultata.

Prebrojeni histogrami su histogrami boja koji su višedimenzionalni i samim time nemaju probleme neprebrojenih histograma. Da bih razulemeli ovaj koncept bolje važno je gledati mapu boja kako kocku, zato što ima tri primarne boje (crvena, zelena, plava) i objekat koji predstavlja boje mora biti tridimenzionalan. Crvena boja je predstavljena kao dimenzija visine, plava boja kao dimenzija širine i zelena boja kao dimenzija dubine.

Kocka ima 256 vrednosti po dimenziji što dozvoljava preko 16 miliona drugačijih boja (16777216). Najpreciznije predstavljanje vektora nastaje kada svaka pozicija odgovara jednoj boji, ali to dovodi do ogromnog vektora gde će skoro sve pozicije biti nule (raštrkan vektor). Varijacija bilo kog piksela pravi totalno drugačiju boju iako je vizuelno to nemoguće uočiti. Ovo dovodi do metode podele prostora boje. Ovaj proces se sastoji od podele svake boje u broj izabran od korisnika i onda prebrojavanje broja piksela u ovom dometu. Perceptivno, prostor boja je podeljen tako da više ne sadrži 256 vrednosti po dimenziji. Svaka dimenzija je isečena i sada formira prostor boja sa manje boja nego ranije kao što je prikazano u slici 4.



Slika 4 – Predstavljanje parčića prostora boja za prebrojene histograme

Ovako sve vrednosti iz istog dometa se smatraju istom bojom. Kada se bira piksel iz originalne slike locira se na novom prostoru boja i njegova nova vrednost se dodaje. Kada se pravi višedimenzionalno prebrojavanje, ovaj prostor boja je smanjen i vektor je manje raštrkan.

Primer: za prebrojavanje za vrednost 8 podelićemo prostor boja od 256 vrednosti u 32 delova, svaki veličine 8 (pozicije 0 do 7). Za piksel 17 19 16 dodacemo 1 u poziciju $(2 \times 32 \times 32) + (2 \times 32) + 2$ pošto 17 spada u domet 2 (16 do 23) kao i 19 i 16.

$$Pozicija = \sum_{x=0}^2 \lfloor \frac{v[x]}{q} \rfloor \left(\frac{256}{q} \right)^{2-x}$$

U ovoj funkciji q predstavlja domet vrednosti za svaku boju, a x poziciju crvene, plave i zelene boje u svakom pikselu. U ovom primeru iskorišćavamo svojstva boja da kreiramo vektor, pošto ovo manje više znači da svaka boja sadrži q vrednosti. Broj prebrojavanja je veoma važan zato što može dovesti do raštrkanog vektora u sličaju da je izabran broj veoma mali ili do pogrešne klasifikacije i konfuzije ako je previše veliki kao što se može videti u tabeli 1.

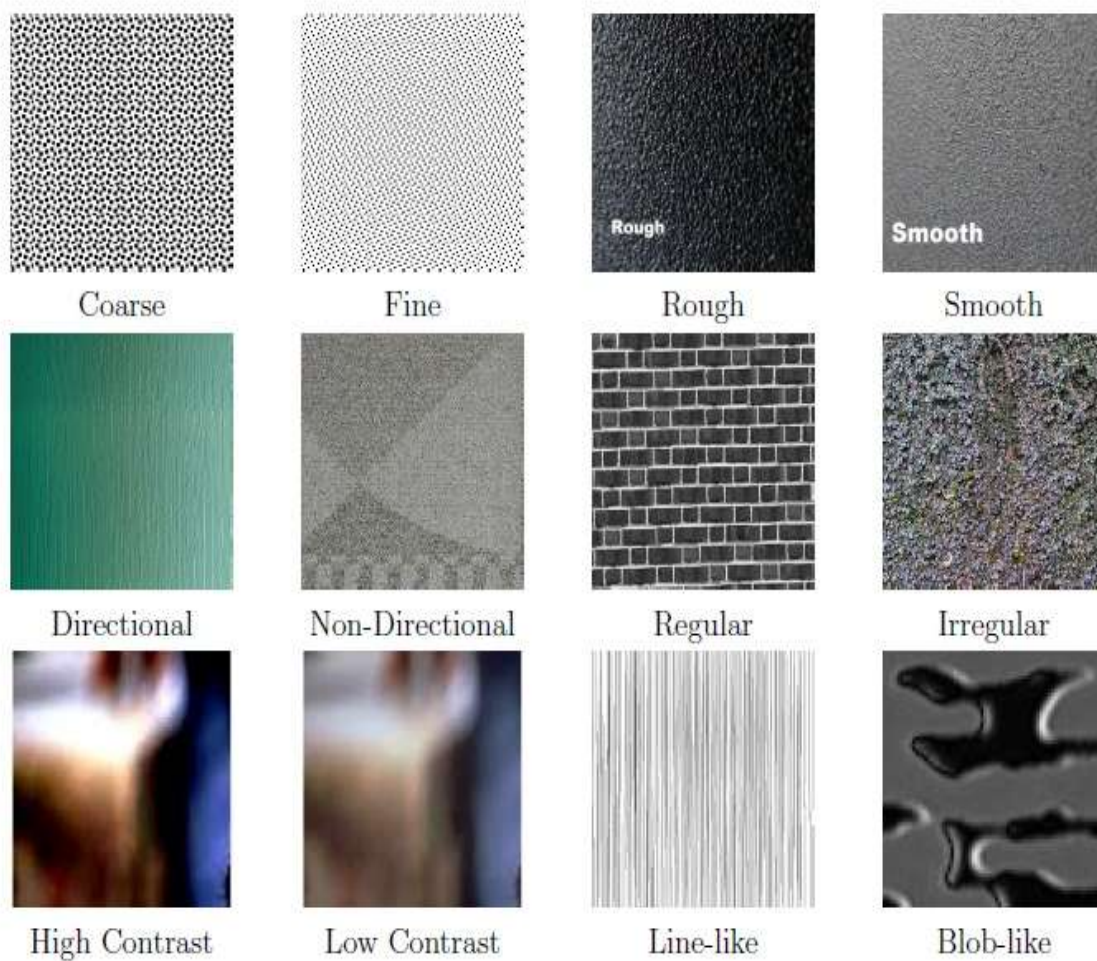
Prebrojavanje	Dužina vektora
1	16777216
8	32768
16	4096
32	512
64	64

Tabela 1 – Dužina vektora pri podeli

Tabela 1 je veoma važna, pošto ne razmatramo samo preciznosti već i vreme računanja. Korisnik želi da mu se upit odgovori što je pre moguće.

7.1.2 Tamura odlike tekstura

Tekstura je ključna komponenta ljudske vizuelene percepcije i isto kao boja je ključna odlika pri pretraživanju kolekcija slika. Teško je definisati teksturu, mada možemo reći da obuhvata neformalne odlike kvaliteta kao što su: hrapavost, glatkoća, gruboća, granularnost, linearnost, usmerenost i pravilnost [26]. Nakon više istraživanja utvrđeno je da su jedine značajne i izračunjive odlike hrapavost, usmerenost i kontrast. Sve Tamura texture se mogu videti u slici 5.

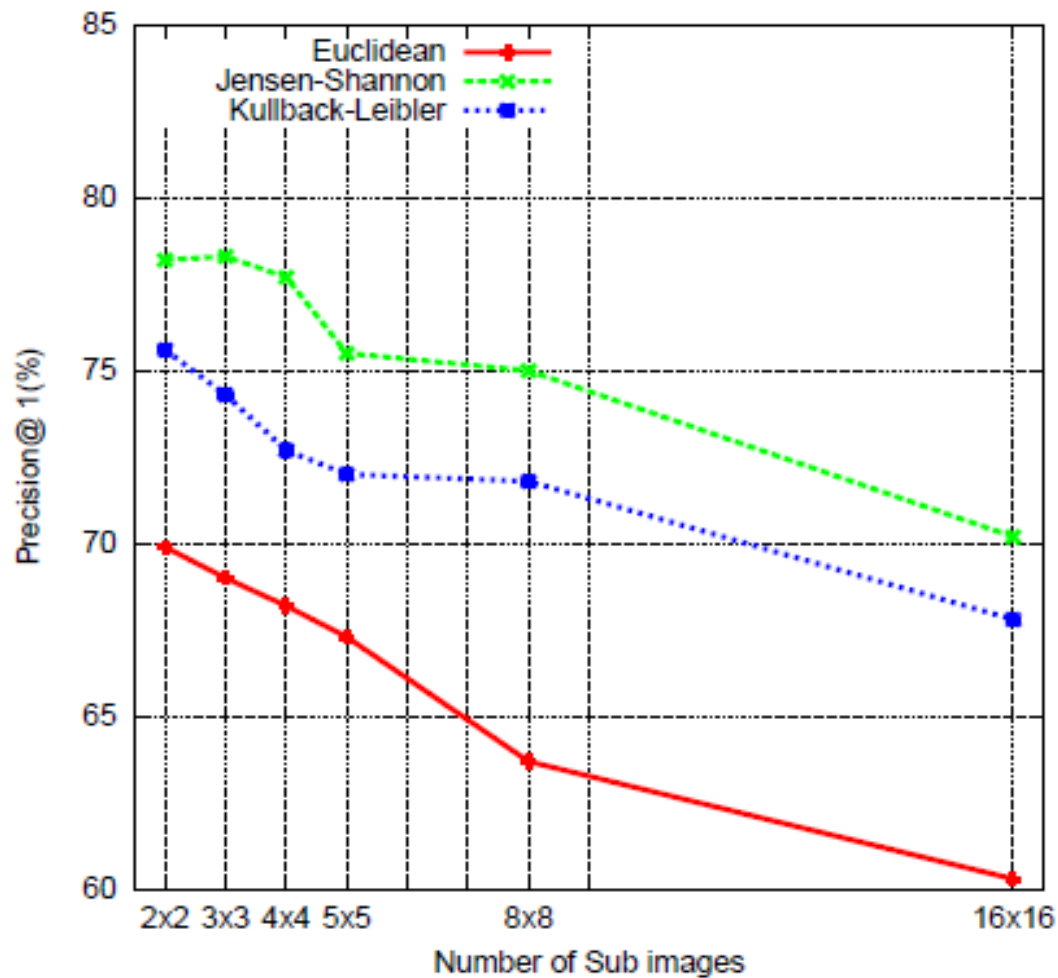


Slika 5 – Primer različitih tekstura

7.1.3 Raspored histograma

Kao poseban slučaj histograma boja, ovo predstavljanje izvlači informacije iz nekih delova slike. To jeste, deli sliku u različite regije i pravi podslike. Kasnije, nezavisan histogram je izvučen iz svake ove slike. Glavni cilj ovog sistema je da koristi informaciju pozicija slika i boja.

Možemo primetiti kako ovaj proces možemo podesiti na mnogo načina. Broj podslika je definitivno veoma važan. Za određenu sliku moramo odlučiti na koliko podslika će biti isečena. Na primer, ako je neka slika podeljena u 9 podslika i posmatramo svaku od njih, možda nećemo primetiti nekakav objekat ako je glavna slika previše zumirana. Na drugu stranu, slika sa više detalja može da prikaže totalno drugačije objekte ili scene u svakoj podslici. Dakle, ovo znači da svakoj slici može da treba drugačiji broj podele zato što ne postoji fiksni broj podslika za sve slike. Vizuelni primer koji objašnjava kako broj podslika utiče na konačnu preciznost se može videti na slici 6.



Slika 6 – Preciznost za raspored histograma koristeći različite mere razdaljine

Prvi pristup ovom metodu će biti deljenje svake slike u različite podslike i izvlačenje histograma boja iz svake od njih, i kreiranje velikog vektora sa svim histogramima boja od slike. Onda, kada znamo broj podslika, upoređićemo ovaj veliki vektor histograma boja sa svim ostalima. Ovo znači da će svaka podslika biti upoređena samo sa onom na istoj poziciji u različitim slikama.

7.1.4 Lokalne odlike

Slično kao i raspored histograma model lokalnih odlika izvlači delimične informacije iz slike, ali u ovom slučaju izvlači se mnogo više detalja. Glavni cilj ove odlike je da poredi različite detalje koji se nalaze u svim slikama.

U ovom slučaju više nije bitna cela slika nego samo male oblasti. Dakle ako se ove male oblasti uglavnom sadrže i u drugim slikama verovatno je da se odnose na istu klasu. Ovo će se desiti u slučajevima kada pravimo sliku istog objekta iz različitih uglova, jačine svetla, pozicije, razdaljine, kamera.

Ovakav tip predstavljanja slike je uobičajen u zadatku prepoznavanja objekata. Prepoznavanje objekata i sistemi za pretraživanje slika bazirani na sadržaju su jako bliske oblasti i za neka pretraživanja, prepoznavanje objekata je jedino izvodljivo rešenje [27]. Ove odlike su deo nepromenljivih odlika. Neka odlika se smatra nepromenljivom u odnosu na neku transformaciju ako se ne menja kada se ta transformacija primeni na sliku. Transformacije na koje se misli su rotacija, skaliranje i preslikavanje.

Proces klasifikacije sa lokalnim odlikama se sadrži iz dva dela: faza treniranja i faza testiranja. U fazi treniranja, lokalne odlike se izvlače iz svih slika za treniranje i dobija se ogroman broj lokalnih odlika. Onda se primenjuje smanjenje dimenzionalnosti od analize glavne komponente (Principal Component Analysis). Iako su rezultati preciznosti za lokalne odlike veći od svih drugih metoda, proces ima dva dela to čini poređenje sporijim.

7.1.5 Lokalni histogrami boja

Slično kao i u pristupu lokalnih odlika, cilj je izabrati deo slike i kasnije izvući histogram boja iz tog dela. To što dobijamo ovde je malo drugačije nego ranije sa lokalnim odlikama. Pošto dobijamo histogram, pozicija svakog piksela i boje nije bitna tako da je poređenje između boja delova slike. Dakle, veličina prozora može biti malo veća pošto se histogram ne menja u odnosu na poziciju, i veličina svakog parčeta (lokalnih odlika) će biti manja. Ranije prozor veličine 11x11 je kreirao 363 vrednosti, dok će se sada vektor kreirati od 64 vrednosti. Prebrojeni histogrami će se koristiti pri izvlačenju glavnih odlika svakog parčeta, i prebrojavanje za vrednost 64 će biti konačna veličina histograma jer su rezultati veoma bliski 32 prebrojavanju i teži mnogo manje.

7.1.6 GIST

GIST predstavljanje je prvi put predloženo u [28]. Ideja je da se napravi malodimenzionalna reprezentacija scene, koja ne zahteva nikakvu segmentaciju. Autori predlažu set perceptualnih dimenzija (prirodnost, otvorenost, hrapavost, proširenost, tvrdoća) koje predstavljaju dominantnu prostornu strukturu scene. Oni pokazuju da te dimenzije mogu da se pouzdano proračunaju. Slika se deli u 4x4 rešetku za koju se izvlače histogrami orijentacije.

8 POVRATNE INFORMACIJE O RELEVANTNOSTI

Povratne informacije o relevantnosti omogućavaju korisniku da interaktivno preradi upit preko specifikacija relevantnih stavki. Uključivanjem korisnika u sistem pretrage možemo dobiti bolje performanse pretrage. To uglavnom izgleda tako što sistem vrati set mogućih rešenja, i korisnik izabere za svaku stavku da li je relevantna ili ne. U ovom slučaju korisnik počinje upit sa određenom slikom, i onda mu je prikazan set slika koje sistem smatra da su relevantne za tu sliku; od ovih slika korisnik označava koje slike su relevantne a koje nisu (možda će neke slike ostati neoznačene) i onda će sistem za pretraživanje preraditi upit koje će dovesti do boljih rezultata za svaku ponavljanje ovog procesa [29].

Ako zamislimo U kao univerzalan skup slika i $C \subseteq U$ kao fiksnu, konačnu kolekciju slika, početna slika za upit koju je korisnik uneo je $q \in U$. Predpostavićemo da korisnik zna neki skup relevantnih slika za sliku upita $R \subseteq U$. Ovaj skup je nepoznat i cilj sistema je da sazna n slika iz njega, među slikama iz C . Interaktivno pretraživanje počinje sa korisničim unosem određene slike $q \in U$. Onda sistem prikazuje inicijalni skup slika $X \subseteq C$ od n slika koje su slične q po pogodnoj meri distance. Ove slike se ocenjuju od strane korisnika koji daje povratne informacije tako što određuje koje slike su relevantne (i implicitno koje nisu). Te povratne informacije se koriste u sistemu za dobjanje novog skupa slika X i ovaj proces se ponavlja dok se korisnik ne zadovolji, što bi značilo da on/ona smatra sve slike u skupu X relevantnim.

Bilo koji korak ovog procesa ćemo označiti kao $F = (Q^+ \cup Q^-) \subseteq C^m$, gde je $m \geq n$ broj slika koje posmatra korisnik, $Q^+ \subseteq R$ su slike koje korisnik smatra relevantnim i $Q^- \subseteq C \setminus R$ su slike koje korisnik ne smatra relevantnim. $C_F = CF$ će biti slike iz kolekcije koje nisu prikazane. Uglavnom se smatra da će početni upit biti u skupu relevantnih slika $q \in Q^+$.

Da bi optimizovali doživljaj korisnika, moramo da maksimalno povećamo verovatnoću da su slike u X relevantne, po F . To jeste, slike u X bi trebalo da su slične slikama u Q^+ i različite od slika u Q .

$$X = \underset{X \in C^n}{argmax} Pr(X \vee C, q, F)$$

Sko uprostimo formulu kao što je objašnjeno u [29], relevantnost slike X može se izračunati kao:

$$q \in \frac{Q^{+d(q,x)^{-1}}}{\sum_{q \in F} d(q,x)^{-1}}$$

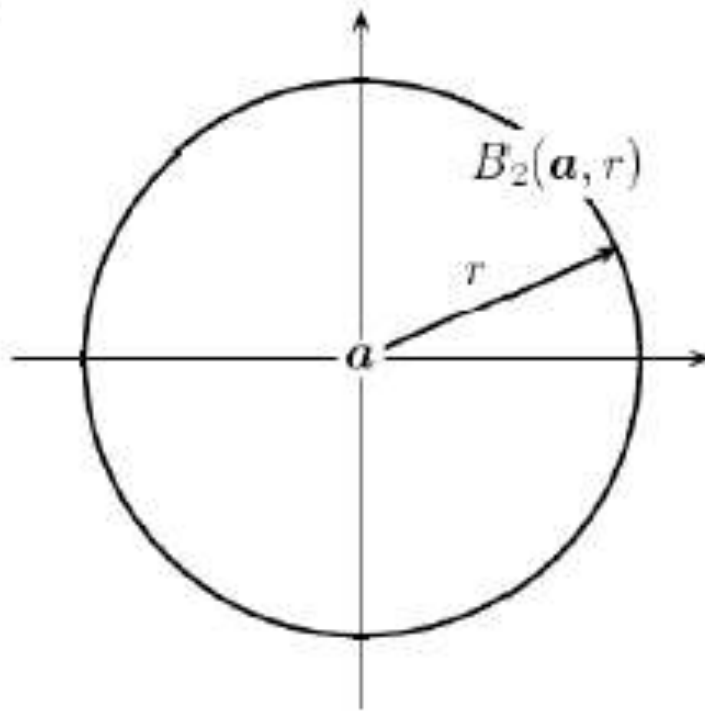
$$\sum Relevance(X) =$$

8.1 Metode razdaljine

Neophodno je da imamo neku meru koja će označavati kvalitet sistema i porediti više slika ili reči i da odredi koje je najbližnja slika ili reč za zadati upit. Pokazaćemo da mere razdaljine igraju bitnu ulogu u klasifikaciji pošto neke mogu biti pogodnije za određene odlike od drugih.

8.1.1 Euklidova razdaljina

Euklidova razdaljina ili L_2 je najviše korišćena distanca i sadrži se iz pravolinijske razdaljne između 2 tačke. Predstavljanje svih tačaka koje imaju istu razdaljinu za određenu tačku u 2 dimenzije je krug, kao što se može videti na slici 7, u 3 dimenzije je sfera itd...



Slika 7 – U dve dimenzije, Euklidova razdaljina ima oblik kruga

$$d_2(x, y) = \sqrt{\sum_a (x_a - y_a)^2}$$

8.1.2 L_1

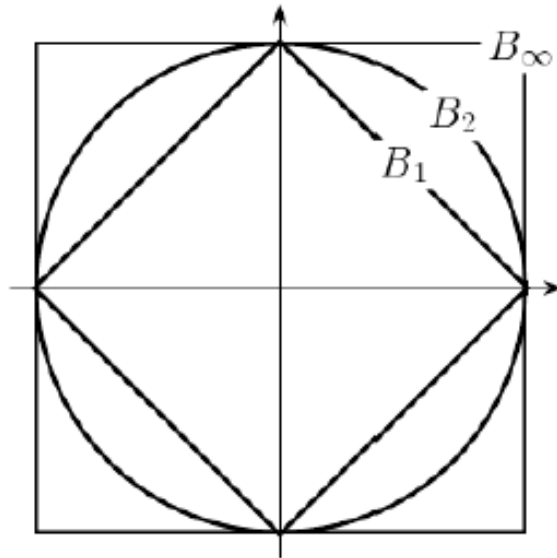
Euklidova razdaljina računa razlike kvadrata, dakle može da zavisi previše od tačaka koje se nalaze daleko jedna od druge. L_1 je više „pravilnija“ pošto samo dodaje razlike u apsolutnoj vrednosti.

$$d_1(x, y) = \sum_d |x_d - y_d|$$

8.1.3 L_∞

L_∞ samo razmatra maksimalnu razdaljinu od 2 tačke u apsolutnoj vrednosti.

$$d_\infty(x, y) = \max_d |x_d - y_d|$$



Slika 8 – Relacije između L_1 , L_2 i L_∞ razdaljina

8.1.4 Dzensen-Šanon

Ovaj metod se uobičajeno koristi u verovatnoći i statistici, i meri sličnost između 2 raspodele. U ovom slučaju koristi se u histogramima radi određivanja nivoa sličnosti. Postoje nekoliko načina implementacije ovog algoritma u slučaju da nema delilac. Dzensen-Šanon i Kulbak-Leibler tehnike će biti pojednostavljene u verzije bez delilaca u nastavku ovog rada.

$$d_{JSD}(H, H') = \sum_{m=1}^M H_m \log \frac{2H_m}{H_m + H'_m} + H'_m \log \frac{2H'_m}{H_m + H'_m}$$

Gde su H i H' histogrami koji se upoređuju a m je broj vrednosti u vektoru.

8.1.5 Kulbak-Leibler

Kulbak-Leibler je nesimetrična verzija Dzensen-Šanon-a.

$$d_{KL}(H, H') = \sum_{m=1}^M H_m \log \frac{2H_m}{H_m + H'_m}$$

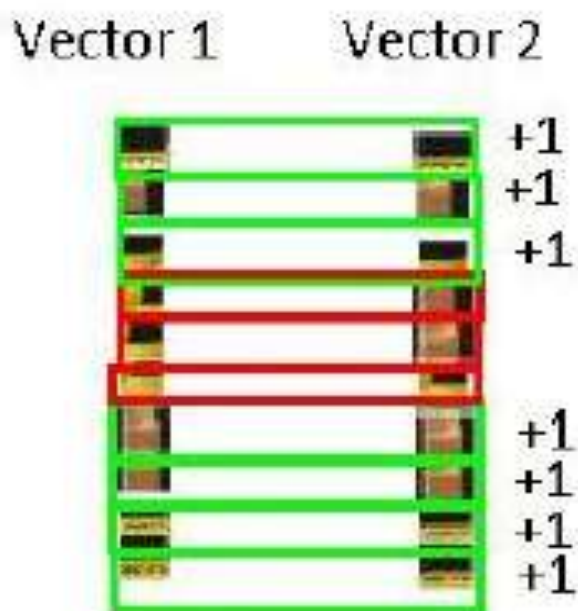
8.1.6 Distorzija

Ovaj tip razdaljine se koristi pri upoređivanju rasporeda histograma. Cilj rasporeda histograma je da upoređi različite male slike na različitim pozicijama koje mogu imati sličan izgled. Razdaljine koje smo razmatrali do sada ne dozvoljavaju poređenje malih oblasti koje se nalaze na različitim pozicijama.

Ovaj metod poređuje svaku podsliku sa podslikama blizu nje i odabira najmanju razdaljinu. Nema smisla upoređivati sve podslike od određene slike sa svim podslikama neke druge slike jer bi to povećalo cenu procesa. Ovo rešenje dozvoljava sličnim objektima ili scenama da se nađu u različitim pozicijama u drugoj slici i u isto vreme da mogu da se klasifikuju u istu klasu. Oba histograma (boja i tekstura) nisu zabeležila nikakvu poziciju piksela kada je poređenje napravljeno u odnosu na boje ili teksture cele slike. Raspored histograma izvlači karakteristike boja iz konkretne oblasti slike i zato je potrebno dozvoliti neke slobode pri poređenju različitih delova slike.

8.1.7 Razdaljina vizuelnih reči

Ova razdaljina je bazirana na metodu praga. Svrha joj je da veoma brzo izračuna razdaljinu između 2 vektora sa dobrim rezultatom. Radi samo za vektore lokalnih odlika pošto nema smisla koristiti je u ostalim metodama. Kada se prag definiše, algoritam uzima oba vektora i poredi svaku poziciju kao u slici 9.



Slika 9 – Primer razdaljine reči korišćene na 2 različita vektora napravljena od različitih lokalnih odlika

Ako su obe pozicije koje se porede veće od definisanog praga algoritam to računa kao 1. Dakle algoritam računa koliko vizuelnih reči su zajedničke između 2 slike. Rezultat će biti pozitivan i veći ako ima više zajedničkih reči. Ali, sistem radi sa distancama tako da je neophodno prevesti ovaj rezultat da bi pretraživanje moglo da radi normalno, tako što ćemo ga napraviti negativnim. Dakle što više zajedničkih parčića, manje razdaljine između njih.

8.2 Metode ocenjivanja

Zamislamo bazu podataka $\{x_i, \dots, x_n\}$ koja je skup slika predstavljenih odlikama. Za upit slike q svaka slika x_i će biti izmerena za određeni metod razdaljine $d(q, x_i)$. Onda se baza podataka sortira opadajuće počinjući sa najmanjom razdaljinom. Za ocenjivanje sistema za pretraživanje slika postoje 2 glavne mere; preciznost P i opoziv R .

$$P = \frac{NR}{TR}$$

$$R = \frac{NR}{TI}$$

Gde je NR broj pronađenih relevantnih slika, TR ukupan broj pronađenih slika i TI ukupan broj relevantnih slika. Preciznost i opoziv su dve širokokorišćene statističke klasifikacije. Preciznost se može videti kao mera tačnosti a opoziv kao mera potpunosti.

Primer: Pri traženju 10 objekata, dobili smo 6 tačna rezultata (iz iste klase) i 4 netačna rezultata. Tražena klasa ima 8 elemenata, tako da 2 tražena objekta se nisu pojavila. Imamo preciznost od 60 posto (6 od 10) i opoziv od 75 posto (6 od 8). Postaje jasno da što je veći opoziv, manja je preciznost (ako želimo da dobijemo sve interesantne rezultate, doćemo i dosta pogrešnih rezultata sa njima)

Vrednosti preciznosti i opoziva se uglavnom predstavljaju u grafikon $R \rightarrow P(R)$ u kojem su sažeti $(R, P(R))$ parovi za različite brojeve dobijenih slika. Najčešće se ovaj grafikon sažima u jednu vrednost preko srednje vrednosti prosečne preciznosti.

Prosečna preciznost PP za upit q je srednja vrednost preciznosti posle izvlačenja svake relevantne stavke:

$$PP(q) = \frac{1}{N_R} \sum_{n=1}^{N_R} P_q(R_n)$$

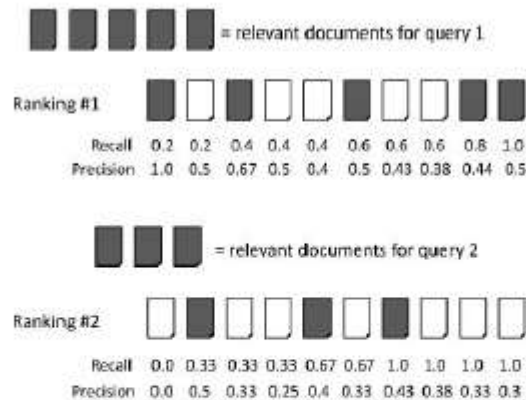
Gde je R_n opoziv posle izvlačenja n -te relevantne slike, a N_R broj relevantnih slika za određen upit. Srednja vrednost prosečne preciznosti $SVPP$ je srednja vrednost prosečne preciznosti svih upita.

$$SVPP(q) = \frac{1}{|Q|} \sum_{q \in Q} PP(q)$$

Drugim rećima, da bi dobili $SVPP$ za određen upit, sve razdaljine do ostalih slika moraju da se poređaju (na primer koristeći algoritam za brzo sortiranje). Kada je to završeno moramo da nađemo sve relevantne slike i da ih podelimo sa pozicijom u kojoj je se svaka od njih nalazi, saberemo sve njih i podelimo sa ukupnim brojem relevantnih slika.

Ako na primer imamo 5 relevantnih slika za određen upit, izračunaćemo preciznost kada su u poređanom vektoru slika prve 5 pozicije relevantne slike.

$$\frac{1}{5} \sum \frac{1}{1} + \frac{2}{2} + \frac{3}{3} + \frac{4}{4} + \frac{5}{5}$$



$$\text{average precision query 1} = (1.0 + 0.67 + 0.5 + 0.44 + 0.5)/5 = 0.62$$

$$\text{average precision query 2} = (0.5 + 0.4 + 0.43)/3 = 0.44$$

$$\text{mean average precision} = (0.62 + 0.44)/2 = 0.53$$

Slika 10 – Primer prosečne preciznosti i srednja vrednost prosečne preciznosti

Glavna prednost korišćenja ove mere je to što sadrži aspekte i preciznosti i opoziva, i osteljiva je na celo rangiranje.

Preciznost@Q je mera preciznosti. Za odedenu sliku upita i broj slika Q, preciznost se dobija deljenjem broja relevantnih slika dobijenih među ukupnim brojem slika od Q.

$$\text{Preciznost@Q} = \frac{1}{|Q|} \sum_{q \in Q} \begin{cases} 1 & \text{ako je najslabija slika relevantna} \\ 0 & \text{inak} \end{cases}$$

U ovom slučaju preciznost je izmerena, mada je veoma često da se mere greške sa $1 - P(Q)$ gde je najčešći Q broj 1, sa ciljem dobijanja najslabije slike za svaki upit.

9 Zaključna razmatranja

U svom istraživanju u ovom radu došli smo do sledećih zaključaka. Oblast pronalaženja informacija se jako unapredila tokom zadnjih 40 godina. Iako se u prvim godinama istraživanja sumljalo u efikasnost jednostavnih statističkih tehnika, one su pokazale kao najefikasnijim do sada. Pretraživanje slika bazirano na sadržaju se pokazalo boljim od pretraživanja slika baziranim na tekstu, ali najbolje rešenje je kombinacija oba sistema. Od svih mera razdaljina Dzensen-Šanon je najbolja za skoro svaku različitu tehniku izvlačenja vizuelnih odlika, a za reči razdaljina reči.

Tehnike razvijene u oblasti pronalaženja informacija su se koristile u mnogo drugih oblasti i doprinele mnogim novim tehnologijama koje se koriste svakodnevno, kao na primer web pretraživači, filteri spam email-ova I upoređivači slika u policiji.

Dalja istraživanje u ovoj oblasti može rešiti razne probleme sa kojima se susreću korisnici u današnjem svetu prepunom informacija, kao i značajno olakšati dostupnost bitnim informacijama. Sa eksponencijalnim rastom količine informacija koja nam je dostupna, pronalaženje informacija će imati sve bitniju ulogu u našim životima.

Literatura

1. Turing, A. (1950). Računanje i inteligencija. *Mind*, 59, 433–460.
2. Vannevar Bush. (1945). Kako možemo misliti. *Atlantic Monthly*, 176:101–108.
3. Gerard Salton, editor. (1971). *The SMART Retrieval System—Experiments in Automatic Document Retrieval*.
4. C. W. Cleverdon. (1967). Cranfield testovi. *Aslib Proceedings*, 19:173–192.
5. D. K. Harman. (1993). Pregled prve TREC konferencije. *Proceedings of the First Text REtrieval Conference (TREC-1)*, pages 1–20.
6. Gerard Salton, A. Wong, and C. S. Yang. (1975). Model vektorskog prostora za pretragu informacija. *Communications of the ACM*, 18(11):613–620.
7. S. E. Robertson. (1977). Rankiranje pomoću modela verovatnoće. *Journal of Documentation*, 33:294–304.
8. M. E. Maron and J. L. Kuhns. (1960). O relevanci, indeksiranju pomoću verovatnoće i pretrazi informacija. *Journal of the ACM*, 7:216–244.
9. W. B. Croft and D. J. Harper. (1979). Korišćenje modela verovatnoće u pretrazi dokumenata bez informacije o relevantnosti. *Journal of Documentation*, 35:285–295.
10. Howard Turtle. (1990). *Inference Networks for Document Retrieval*. Ph.D. thesis, Department of Computer Science, University of Massachusetts, Amherst, MA 01003.
11. David Hull. (1996). Algoritmi kalemljenja – istraživanje slučaja za detaljno procenjivanje. *Journal of the American Society for Information Science*, 47(1):70–84.
12. J. L. Fagan. (1989). Efektivnost indeksiranja statističkih i lingvističkih fraza za pretragu dokumenata. *Journal of the American Society for Information Science*, 40(2):115–139.
13. Robertson, S. E. and Sparck Jones, K. (1976). Ocenjivanje težina relevantnosti termina. *J. American Society for Information Science*, 27, 129–146.
14. Gerard Salton and M. J. McGill. (1983). *Introduction to Modern Information Retrieval*.
15. K. Sparck Jones. (1971). *Automatic Keyword Classification for Information Retrieval*.
16. J. J. Rocchio. (1971). Relevantnost povratnih informacija u pretrazi informacija. Gerard Salton, editor, *The SMART Retrieval System—Experiments in Automatic Document Processing*, pages 313–323.
17. Chris Buckley, James Allan, Gerard Salton, and Amit Singhal. (1995). Automatsko proširivanje upita pomoću SMART-a. *Proceedings of the Third Text REtrieval Conference (TREC-3)*, pages 69–80.
18. Banko, M., Brill, E., Dumais, S. T., and Lin, J. (2002). AskMSR: Odgovaranje na pitanja pomoću web-a. In *Proc. AAAI Spring Symposium on Mining Answers from Texts and Knowledge Bases*, pp. 7–9.
19. Stuart Russell and Peter Norvig. (2009). *Artificial Intelligence: A Modern Approach*.
20. K. S. Candan and M. L. Sapino. (2010). *Data Management for Multimedia Retrieval*: Cambridge University Press
21. R. Datta, et al. (2008). Pretraživanje slika: Ideje, uticaji, i trendovi novog veka, *ACM Computing Surveys (CSUR)*, vol. 40, p. 5.
22. J. Jeon, et al. (2003). Automatska anotacija slika i pretraživanje uz pomoć raznomedijskih modela relevantnosti, pp. 119-126.
23. I. Jolliffe. (2002). Analiza glavnog komponenta , *Encyclopedia of Statistics in Behavioral Science*, vol. 3, pp. 1580-1584.

24. D. Mehta, E.S.V.N.L.S. Diwakar, and C.V. Jawahar. (2003). *A rule-based approach to image retrieval*, vol 2, pp. 586-590.
25. A. P. de Vries and T. Westerveld. (2004). Poređenje diskretnog i neprekidnog modela za pretraživanje slika, *Proceedings of the International Conference on Image Processing (ICIP)*.
26. Winston H. Hsu. (2007). Tekstura i oblici za pretraživanje slika – analiza multimedije i indeksiranje.
27. Thomas Deselaers, Daniel Keysers, and Hermann Ney. (2004). Odlike za pretraživanje slika – kvantitativno poređenje. *DAGM, vol 3175 of LNCS*, pp. 228-236.
28. Aude Oliva and Antonio Torralba. (2001). Modelovanje oblika scene: holistička reprezentacija prostorne koverta. *International Journal of Computer Vision*, pp. 145-175.
29. Roberto Paredes, Thomas Deselaers, and Enrique Vidal. (2008). Model verovatnoće korisničkih povratnih informacija o relevantnosti za pretraživanje slika. *MLMI*, pp. 260-271.