Tehnologije i platforme za upravljanje elektronskim sadržajima i dokumentima

1. Objasniti pojam dokumenta, papirnog dokumenta i digitalnog dokumenta.

Pojam dokumenta obuhvata tradicionalne papirne dokumente i računarski obrađene informacije kojima se rukuje kao osnovnom jedinicom obrade

Papirni dokument čovek(službenik) čita, obrađuje i arhivira, a najčešće ga popunjava klijent upotrebom hemijske olovke. Papirni dokument se može kreirati na praznom papiru pisanjem ili štampanjem.

Digitalni dokument je zapravo računarski obrađena informacija kojom se rukuje kao osnovnom jedinicom obrade. To su: tekstualni digitalni dokumenti(tekstualni opisi ili poruke), grafički dokumenti(slike, crteži, dijagrami, grafikoni), strukturirani dokumenti(HTML i XML+XLink dokumenti), mediji sa vremenskom dimenzijom(zvuk i video), kompozitni multimedijalni dokumenti(sastavljeni od teksta, slike, zvuka, ili videa).

2. Šta su to metapodaci i navesti nekoliko primera metapodataka?

Metapodaci su podaci o dokumentu, odnosno podaci o podacima. Svrha je organizovanje kolekcije sadržaja (dokumenata), klasifikacija dokumenata, pretraživanje dokumenata. Primeri:

Za tekstualni digitalni dokument: autor, naslov, datum nastanka, ključne reči.

Za digitalnu fotografiju: autor, datum i vreme fotografisanja, mesto fotografisanja, podešavanje aparata, objekti prikazani na slici

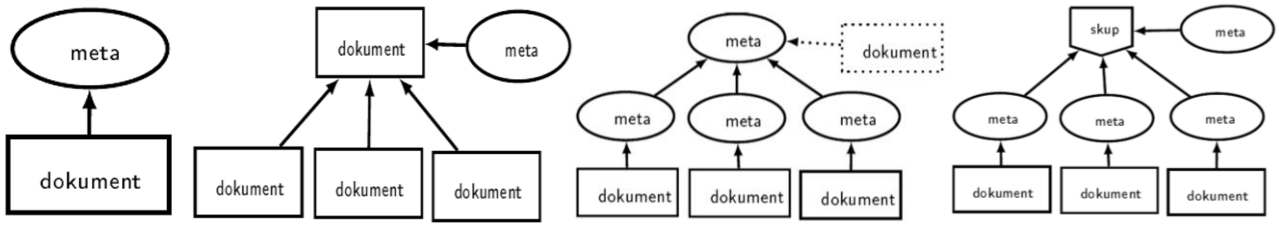
3. Kakve sve veze između dokumenta i metapodataka mogu postojati?

Pojedinačni dokument: Elementarni oblik nosioca informacija. Ima pridružene metapodatke koji opisuju njegov sadržaj ili druge karakteristike(metapodaci: podaci o podacima).

Složeni dokument: Kompozicija više dokumenata različitih tipova (npr. tehnička specifikacija koja se sastoji od tekstualnih fragmenata, crteža, i dijagrama). Metapodaci se pridružuju složenom dokumentu kao celini.

Agregacija dokumenata: Skup samostalnih dokumenata, svakog sa svojim metapodacima. Agregacija poseduje sopstvene metapodatke. Može, a ne mora, da poseduje poseban sopstveni dokument.

Skup dokumenata: Poseduje svoje metapodatke koji opisuju svrhu skupa, kao i sadržanih dokumenata u skupu.



Pojedinačni dokument Složeni dokument Agregacija dokumenata Skup dokumenata

Aktivna veza: stanje u kome deo sadržaja jednog dokumenta biva preuzet ili na neki drugi način zavisi od sadržaja drugog dokumenta, izmenom drugog dokumenta menja se i prvi.

4. Koje su faze životnog ciklusa dokumenta?

Inicijalizacija: Prva faza koja predstavlja formiranje podataka potrebnih za kasniju pripremu, ali ne obuhvata pripremu i utvrđivanje sadržaja. Rezultat ove faze je okvir u kome se dalje priprema dokument.

Priprema: Proizvodnja sadržaja dokumenta sve do trenutka uspostavljanja. Počinje nakon inicijalizacije. Metapodaci koji se dodaju u ovoj fazi bi mogli da sadrže(nivo razvoja dokumenta, ključne reči, rezime ili apstrakt, izvor dokumenta).

Uspostavljanje: Pre korišćenja dokument se obično odobrava za potrebe obezbeđivanja kvaliteta i po pravilu se primenjuje na sve verzije. Pravila za odobravanje se definišu na nivou poslovnog procesa, klase dokumenta i pojedinačnog dokumenta.

Korišćenje: Dokumenti su, sa metapodacima, dostupni za korišćenje. Metapodaci se koriste za pretraživanje i informisanje o dokumentima i njihovim verzijama. U metapodatke se mogu dodati komentari/iskustva korisnika o korišćenju dokumenta.

Revizija: Faza koja je opciona i ponavljajuća, i koja predstavlja promenu sadržaja ili promenu namene dokumenta. Nakon ove faze se obično ponovo vrši uspostavljanje, odnosno odobravanje pre početka korišćenja nove verzije. U jednom trenutku može se koristiti više verzija sve dok ispunjavaju svoju namenu.

Arhiviranje: Predstavlja premeštanje dokumenata (verzija, metapodataka) u kompaktniju nepromenljivu formu. Mora da ispuni ugovorne/zakonske obaveze (npr. rok čuvanja). Potrebno je obezbediti kontrolisani pristup arhivi i mogućnost reprodukcije dokumenata. Potrebno je sprečiti mogućnost izmena dokumenata koji su arhivirani. Arhiva je baza znanja zbog čega je potrebno obezbediti pretraživanje arhive digitalnih dokumenata. Poželjno je koristiti stabilne, nepromenljive formate podataka.

Uklanjanje; Dokument se može ukloniti nakon isteka perioda za obavezno arhiviranje. Uklanjanje sadržaja i metapodataka ne mora biti istovremeno(dok se drugi dokument ili verzija referišu na dati dokument trebalo bi čuvati metapodatke). Rezultuje nepovratnim gubitkom podataka, dokumenata i relacija sa drugim dokumentima.

5. Objasniti životnu fazu dokumenta korišćenje.

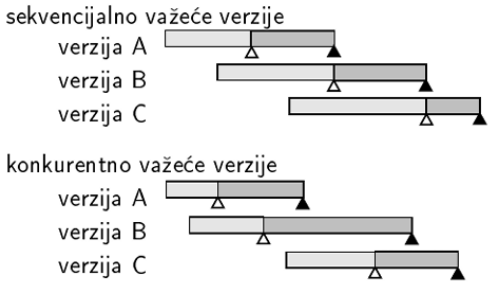
↑

6. Objasniti životnu fazu dokumenta arhiviranje.

↑

7. Objasniti pojmove upravljanje verzijama dokumenata, sekvencijalno i konkurento važenje verzija.

Ako dokument ima više verzija moramo imati podršku za upravljanje verzijama. Za svaku verziju postoji period formiranja kada se verzija formira i period važenja kada se verzija smatra važećom. Važenje verzija se može organizovati sekvencijalno ili konkurentno.



Sekvencijalno važenje: podrazumeva da je poslednja verzija dokumenta jedina važeća. Nova verzija uvek preuzima važenje od prethodne verzije i podržava sve namene svih prethodnih verzija. Odnos zamenjuje/zamenjen navodi se i u metapodacima odgovarajućih verzija.

Konkurentno važenje: podrazumeva da više različitih verzija može biti operativno u jednom trenutku. Nova verzija ne zamenjuje automatski prethodnu u smislu važenja, odnosno verzija ostaje važeća sve do eksplicitnog povlačenja verzije. Povlačenje verzije predstavlja izmene u metapodacima, ali ne i u sadržaju dokumenta.

8. Koja je osnovna namena sistema za upravljanje dokumentima?

Sistem za upravljanje dokumentima: Sistem namenjen praćenju i skladištenju digitalnih dokumenata. Za dokumente koji sadrže nestrukturirane ili slabo strukturirane tekstove i koji nisu pogodni za čuvanje u relacionim bazama podataka.

9. Koje su funkcije sistema za upravljanje dokumentima?

Skladištenje dokumenata: je osnovna namena ovih sistema i postoji dva modela za organizaciju kolekcije dokumenata: 1. centralizovani(jednostavniji, brži, ograničen memorijskim kapacitet),
2. distribuirani(mreža tvrdih diskova na različitim računarima, zavisan od mrežnog protoka(LAN), korisnik ne treba da bude svestan organizacije kolekcije dokumenata i gde se skladišti i treba da ima jedinstven pogled na kolekciju dokumenata).

Katalogizacija: Predstavlja sistematično sređivanje liste zapisa. Ova funkcionalnost je vrlo bitna zbog mogućnosti pretrage i pregleda kolekcije dokumenata.form Može je obavljati jedan ili više učesnika (kolaboracija). Formati metapodataka MARC 21, Dublin Core i ETD-MS.

Pretraživanje: Funkcionalnost sistema koja nam omogućuje pristup znanju koje se nalazi u velikim kolekcijama(http://archive.org/, http://books.google.com/, http://arxiv.org/, WWW) digitalnih dokumenata. Forma za pretragu. Udaljena pretraga iz drugog sistema. Ovim se bavi oblast pronalaženje informacija.

Zaštita podataka: Sprečavanje neautorizovanog preuzimanja i izmene digitalnih dokumenata. Mehanizmi za autentifikaciju i autorizaciju korisnika. Vrednost ovih sistema je u podacima i oni moraju biti zaštićeni. Ako su neki dokumenti otvorenog pristupa (eng. open access) moraju biti rešena autorska i druga pravna pitanja. Može se vezati licenca koja definiše prava korišćenja.

Oporavak od katastrofe: U slučaju katastrofa(požar, poplava, električni udari, fizička oštećenja uređaja...) u kojima se uništavaju sadržaji na tvrdim diskovima, potrebno je da sistem ima implementirane mehanizme za backup(dnevni, nedeljni, mesečni, čuva se na fizički udaljenom uređaju) i recovery(obavlja samo u slučaju potrebe, ali se mora dobro testirati da se ne bi ispostavilo da nakon pretrpljene katastrofe backup ili recovery mehanizam ima propusta).

Arhiviranje: Omogućuje arhiviranje nekog dokumenta koji više nije u upotrebi, odnosno omogućuje formiranje arhive. Mora se omogućiti pregled arhive kao i pretraga arhiviranih dokumenata. Arhiva je baza znanja.

Distribucija: Funkcionalnost koja nam obezbeđuje mehanizme za distribuiranje digitalnih dokumenata. Bavi se pitanjima kako obavestiti javnost da je neki dokument nastao i kako obavestiti javnost da je neki dokument promenjen. Ako za neki digitalni dokument niko ne zna da postoji, to je isto kao i da ne postoji.

Upravljanje poslovnim procesima: Poslovni procesi nastajanja nekog dokumenta. Obezbeđivanje saradnje korisnika.

10. Opisati Dublin Core format metapodataka.

Dublin Core: Standard za reprezentaciju metapodataka. Jedan od najčešće korišćenih formata za opis i razmenu metapodataka između informacionih sistema koji mogu biti i različitih vrsta. Definiše osnovni skup elemenata i

pripadajućih atributa za opis resursa, ali se ovaj skup može i proširivati. Ima dva nivoa:
1. Simple Dublin Core: 15 elemenata za opisivanje metapodataka(title, creator, subject, description, publisher, contributor, date, type, format, identifier, source, language, relation, coverage, rights).
2.Qualified Dublin Core: dodatna 3 elementa i mehanizam rafinacije metapodataka(audience, provenance, rightsHolder, rafinacija elemenata - mogu se dodatno opisati neki elementi).
Dublin Core zapis (record):

```
<record>
  <dc:title>Na Drini ćuprija</dc:title>
  <dc:creator>Ivo Andrić</dc:creator>
  <dc:contributor>Petar Džadžić</dc:contributor>
  <dc:publisher>Srpska književna zadruga</dc:publisher>
  <dc:date>1971-01-01</dc:date>
  <dc:language>sr</dc:language>
  <dc:identifier>ISBN:0140177388</dc:identifier>
</record>
```

11. Šta su protokoli za razmenu podataka?

Standardizovani protokoli u oblasti upravljanja digitalnim dokumetima se mogu podeliti u dve vrste i jedna od njih je protokol za razmenu podataka između dva sistema. Mogu se razmenjivati i dokumenti zajedno sa metapodacima ali najčešće se razmenjuju samo metapodaci i među tim metapodacima nalazi se i metapodatak koji sadrži URL ka digitalnom dokumentu koji ostaje u svom izvornom repozitorijumu(pravima pristupa upravlja izvorni repo a omogućuje se vidljivost i lakše pronalaženje). Njegov predstavnik je OAI PMH protokol.

12. Koje su osnovne karakteristike OAI-PMH protokola?

OAI-PMH: Protokol koji omogućuje razmenu podataka između sistema. Postoje dve klase učesnika:
1. Data Provider- serverska strana protokola koja daje metapodatke iz svog repozitorijuma
2. Service Provider- klijentska strana protokola koja preuzima podatke i inicira komunikaciju, tj. izdaje zahteve na koje Data Provider odgovara. Metapodatke koristi za pružanje različitih vrsta servisa: pretraga, pregledanje, izveštavanje,

Definiše šest vrsta zahteva: Identify, GetRecord, ListMetadataFormats, ListRecords, ListIdentifiers, ListSets. Baziran je na HTTP.

13. Šta su protokoli za udaljeno pretraživanje?

Standardizovani protokoli u oblasti upravljanja digitalnim dokumetima se mogu podeliti u dve vrste i druga od njih je protokol za pretraživanje udaljene kolekcije, odnosno kolekcije koja se ne nalazi na istom računaru kao i program putem kojeg se zadaje upit. Omogućava direktno pretraživanje udaljenih kolekcija bez potrebe da se podaci preuzimaju i čuvaju na centralnom mestu. Ima nekoliko nedostataka, kao što su performanse, otkazivanje udaljenog servera, brzina mrežnog protoka... Njegov predstavnik je Z39.50 i njegov naslednik SRU protokol.

14. Koje su osnovne karakteristike Z39.50 i SRU protokola?

Z39.50: Binarni klijent-server protokol koji koristi transportni sloj Interneta TCP/IP za komunikaciju klijenta i servera i ne zavisi od platforme. Osnovna funkcionalnost protokola je da uspostavi komunikaciju između klijentske i serverske aplikacije, zatim da izvrši zahtev za pretraživanje i na kraju da vrati formatiranu listu rezultata pretrage. Omogućuje bogat upitni jezik koji omogućava korišćenje bulovih izraza, skraćivanje reči, kao i izbor naprednih opcija za pretragu.

SRU (Search/Retrieve via URL): XML orijentisani naslednik Z39.50 protokola. Protokol za pretraživanje i dohvat informacija putem URL-a. Za transport podataka koristi XML dokumente, a za specifikaciju upita koristi se CQL (Contextual Query Language). Ključni napredak je što se za komunikaciju u koriste otvoreni i široko prihvaćeni standardi.

15. Čime se bavi oblast pronalaženja informacija (information retrieval)?

Pronalaženje informacija: Oblast koja se bavi tehnikama za reprezentaciju, skladištenje, organizaciju, pristup i pronalaženje informacija. Ove tehnike treba da obezbede pronalaženje materijala (najčešće dokumenata) nestrukturirane prirode (najčešće tekstualnih) u okviru velike kolekcije koji zadovoljava korisnikove potrebe za informacijama. Sistemi za pronalaženje informacija najčešće imaju odvojene procese indeksiranja i pretraživanja.

16. Koja je razlika između pronalaženja podataka (data retrieval) i pronalaženja informacija (information retrieval)?

Data retrieval: Bavi se pronalaženjem podataka koji zadovoljavaju precizno definisan kriterijum pri čemu se očekuje od korisnika da poznaje strukturu podataka u bazi koju pretražuje. Podaci u kolekciji koja se pretražuje imaju dobro definisanu strukturu i semantiku. Određuje samo koji zapisi u kolekciji sadrže ključne reči koje je korisnik naveo u upitu, što često dovodi do toga da korisnik ne može da pronađe ono što želi.

Information retrieval: Od korisnika ne očekuje da poznaje kolekciju koju pretražuje, korisnika interesuju informacije o nekoj temi, a ne podaci koji zadovoljavaju upit. Podrazumeva se nepreciznost u zadavanju upita, kao i nepreciznost u listi pronađenih rezultata, odnosno u listi rezultata na postavljeni upit se mogu pojaviti dokumenti koji ne zadovoljavaju korisnikovu informacionu potrebu.

17. Kako se razvijala oblast pronalaženja informacija?

Organizacijom informacija za kasnije pronalaženje i upotrebu ljudi se bave poslednjih 4.000 godina. Za inicijalni razvoj ove oblasti najzaslužnija je oblast bibliotekarstva, a oblast je doživela svoju ekspanziju razvojem veba. Prvi računarski podržani sistemi za pretragu su se pojavili u bibliotekama, a danas ih imamo svuda uključujući i pretraživače veba.

18. Kakve arhitekture mogu imati sistemi za pretragu?

Sa stanovišta arhitekture sistemi za pronalaženje informacija se dele na: centralizovane(podaci i resursi za pretragu smešteni na jednom centralnom serveru ili grupi servera) i distribuirane(gde su podaci i resursi za pretragu raspodeljeni preko više lokacija ili servera, često geografskih različitih).

19. Koje vrste sadržaja mogu biti pretraživane putem sistema za pretragu?

Pretraga tekstualnih sadržaja: nestrukturiranih sadržaja i strukturiranih tekstualnih sadržaja koji iako imaju strukturu u nekim poljima svoje strukture imaju velike količine tekstova,
Pretraga linkovanih tekstualnih sadržaja (pretraga veba),
Pretraga multimedijalnih sadržaja: uključuje sliku, zvuk, video,
Pretraga ostalih vrsta sadržaja: kolekcija programskih izvornih kodova, kolekcija 3D objekata, itd...

20. Koji modeli za pretraživanje se koriste u sistemima za pretragu?

Klasični modeli: Bulov model, Vektorski model, Probabilistički model.
Alternativni modeli: Prošireni Bulov model, Fuzzy model, Model neuronske mreže, Jezički model.

21. Koja je razlika između terma i tokena?

Tekstualni digitalni dokumenti se sastoje od reči koje predstavljaju određeni broj znakova u tekstu. Instanca reči koja se pojavljuje u tekstu zove se token u okviru oblasti pronalaženja informacija. Jedna klasa ekvivalencije reči se zove term.

Razlika između tokena i terma: primer - U julu ću ići na odmor u Grčku, na ostrvo Krit.

Prethodna rečenica ima 11 tokena. U srpskom jeziku se reči razdvajaju znacima interpukcije ili belim znacima (white space characters). Broj termova zavisi od pretprocesiranja teksta. Ako prebacujemo sva slova u mala onda su prvi token "U" i sedmi token "u" zapravo isti term. U svakom slučaju peti i deveti token "na" su jedan term.

22. Šta je tokenizacija i koji problemi postoje u ovoj fazi pretprocesiranja?

Tokenizacija teksta je izdvajanje tokena (reči) iz tekstualnog digitalnog dokumenta. Može biti vrlo složena, posebno za jezike koji nemaju jasno definisano razdvajanje reči. Svaki token je kandidat za term.
Problemi: fraze (naučno-istraživačkog, eUprava, e-Uprava, Novi Sad, baza podataka), datumi i brojevi telefona, kinesko pismo (nema belih znakova, neki znakovi nisu jednoznačni), složenice(Nemački: Lebensversicherungsgesellschaftsangestellter = leben + versicherung + gesellschaft + angestellter), japanski jezik(koristi 4 različita alfabeta), arapsko pismo (piše se sa desna u levo a brojevi sa leva u desno).

23. Zašto se vrši „normalizacija" reči?

Normalizacija se vrši da bi sveli termove u indeksiranom tekstu i u upitima u isti oblik. Na primer, želimo da izjednačimo U.S.A. i USA, window, Window, windows, i Windows, Akcenti: résumé vs. resume, ćevapčići vs. Cevapcici, Umlauti: Universität vs. Universitaet

24. Šta je to steming?

Steming je grubi heuristički proces koji odseca krajeve reči sa ciljem da postigne rezultat što sličniji onome koji postiže pravilna lematizacija bazirana na lingvističkom znanju. Steming je deo procesa normalizacije teksta, često se očekuje da su pre steminga prebačena velika u mala slova, izbačene stop reči... Dobijeni niz znakova se zove stem koji ne mora biti reč koja postoji u jeziku, ne mora koren početne reči.

25. Šta je to lematizacija?

Lematizacija: Redukovanje raznih gramatičkih oblika na baznu formu, tj. podrazumeva pravilnu redukciju na osnovni rečnički oblik (eng. lemu).
Primeri: bih, bi, bismo, biste, biše = biti
automobil, automobili, automobila, automobilu = automobil
voleo bih da kupim automobil = voleo biti da kupiti automobil

26. Objasniti Bulov model pretraživanja.

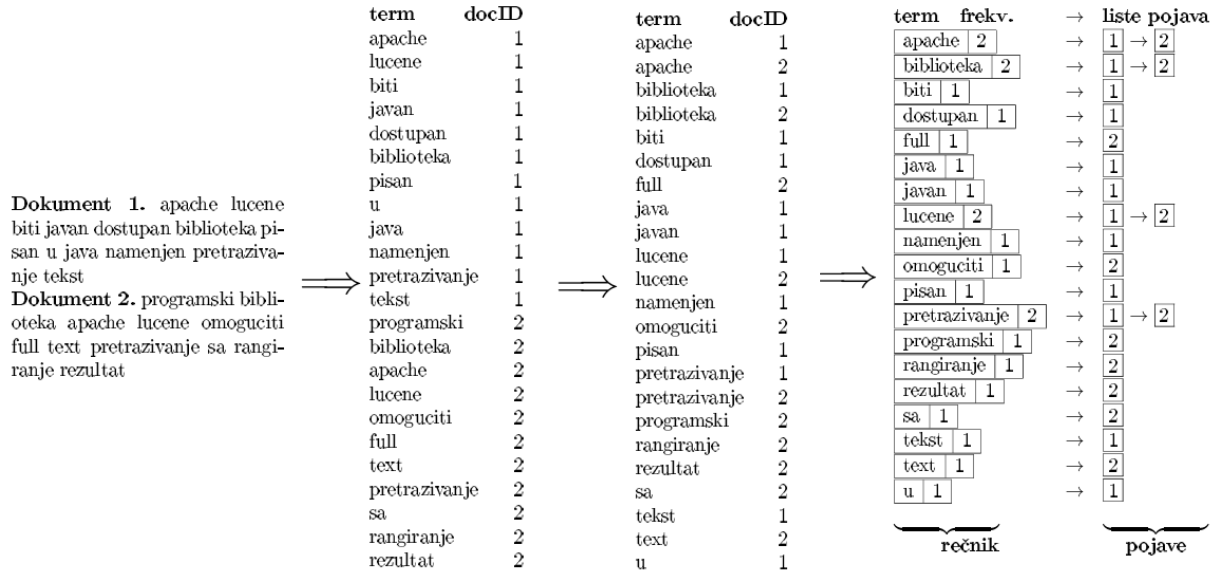
Bulov model pretraživanja je klasični model pretraživanja koji je prvi bio široko prihvaćen u sistemima za pretraživanje. Zasnovan je na teoriji skupova i Bulovoj algebri. Mana mu je što nema mogućnost rangiranja rezultata. Dokument je ili relevantan ili nije, nepostoji parcijalno poklapanje upita i dokumenta.

27. Šta je to invertovani indeks i kako se kreira?

Invertovani indeks predstavlja strukturu podataka koja mapira reči i brojeve iz sadržaja dokumenata na dokumente u kojima se javljaju, a u boljoj varijanti i na lokaciju na kojoj se te reči i brojevi javljaju u dokumentima.

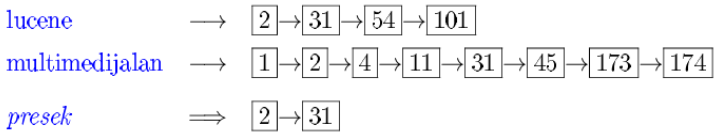
Kreiranje: 1. prikupljanje dokumenata koje treba indeksirati
2. tokenizacija teksta- pretvaranje svakog dokumenta u listu tokena
3. pretprocesiranje teksta- formiranje liste normalizovanih tokena, tj. termova koji će biti u rečniku
4. indeksiranje dokumenata- formiranje invertovanog indeksa koji ima rečnik i pojave. Ovaj korak zahteva da se za svaki term dobijen na kraju trećeg koraka kreira lista pojava ovog terma u dokumentima koji su u kolekciji. Zatim je

porebno sortirati prethodno prikazane pojave u dokumentima. Nakon sortiranja moguće je kreirati liste pojava za svaki term i izračunati frekvencije pojavljivanja. Na kraju je potrebno podeliti rezultat u fajl za rečnik i fajl za pojave i na taj način je kreiran invertovani indeks.



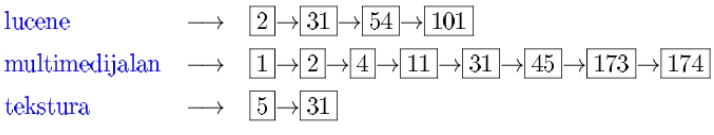
28. Objasniti procesiranje upita kod Bulovog modela.

Razmatramo jednostavan konjunktivni upit dva tokena Lucene and multimedijalnih.
Algoritam za pronalaženje svih relevantnih dokumenata za ovaj upit pomoću invertovanog indeksa je sledeći:
1 pretprocesiranje upita nakon čega se dobija konjuktivni upit dva terma: lucene and multimedijalan,
2 pronalaženje terma lucene u rečniku termova,
3 učitavanje liste pojava ovog terma iz fajla sa pojavama,
4 pronalaženje terma multimedijalan u rečniku termova,
5 učitavanje liste pojava ovog terma iz fajla sa pojavama,
6 izračunavanje preseka ove dve liste pojava,



7 vraćanje rezultata korisniku - vraćanje dokumenata koji se nalazu u prethodno izračunatom preseku

Optimizacija ovog procesiranja je da se upit obradi u rastućem redosledu frekvencije termova, odnosno da se liste pojava sortiraju od najmanje do najveće i da se počne obrada upita od najkraće liste pojava.



Prvo će se sortirati liste pojava i u tom redosledu uraditi Bulova operacija and između njih. U ovom primeru: prvo tekstura and lucene, potom dobijeni rezultat and multimedijalan.

29. Šta su pointeri za preskakanje?

Pointeri za preskakanje se koriste kako bi se ubrzao algoritam za odgovor na korisnikov upit kod Bulovog modela pretraživanja. Omogućavaju preskakanje pojava koje svakako neće biti u rezultatu.

30. Šta se može koristiti ako je potrebno podržati upite fraze?

Upiti fraze se koriste kada korisnici sistema za pretraživanje imaju potrebu da pronađu sadržaje koji sadrže određenu npr. "Fakultet tehničkih nauka".

Za odgovor na upite fraze koriste se ili dvorečni indeksi ili pozicioni indeksi.

31. Šta je to dvorečni indeks?

Dvorečni indeksi pored termova indeksiraju i svaki susedni par reči u tekstu kao frazu. Na primer, "sive markirane pantalone" će dati dva para reči: "sive markirane" i "markirane pantalone". Pored standardnih termova ("sive", "markirane", "pantalone"), svaki od parova reči se tretira kao termin u rečniku.

Upotrebom ovako kreiranog indeksa lako je odgovoriti na dvorečne upite fraze "sive markirane" i "markirane pantalone", ali nije lako odgovoriti na duže upite fraze: "sive markirane pantalone". Ova fraza može da se prikaže kao: "sive markirane" AND "markirane pantalone".

32. Šta je to pozicioni indeks?

Pozicioni indeksi su dobra alternativa za dvorečne indekse. Omogućuju odgovore na upite fraze proizvoljne dužine. Za jedan termin u nepozicionom indeksu je vezana lista pojava ovog termina u dokumentima, pri čemu je svaka pojava docID - identifikator dokumenta u kolekciji koji sadrži termin. Kod pozicionog indeksa svaka pojava je docID i lista pozicija. Pozicioni indeks se može koristiti i za blizinsku pretragu (pronađi sve dokumente koji sadrže pretraga i metapodatak na rastojanju od najviše tri reči).

Na primer, imamo upit to_1 be_2 or_3 not_4 to_5 be_6 i pozicioni indeks

to, 993427:
⟨ 1, 6: ⟨7, 18, 33, 72, 86, 231⟩;
2, 5: ⟨1, 17, 74, 222, 255⟩;
4, 5: ⟨8, 16, 190, 429, 433⟩;
5, 2: ⟨363, 367⟩;
7, 3: ⟨13, 23, 191⟩; . . . ⟩

be, 178239:
⟨ 1, 2: ⟨17, 25⟩;
4, 5: ⟨17, 191, 291, 430, 434⟩;
5, 3: ⟨14, 19, 101⟩; . . . ⟩

Document sa identifikatorom (*docID*) 4 je pogodak!

33. Objasniti Vektorski model pretraživanja.

Vektorski model omogućuje parcijalno poklapanje upita i odgovora, a samim tim omogućuje i rangiranje rezultata. Dokumente, kao i upit, predstavalja u vektorskom prostoru velike dimenzionalnosti. Koliko ima termova u rečniku toliko ima dimenzija u vektorskom prostoru. Za određivanje ovih težina najčešće se koristi tf-idf mera koja uzima u obzir broj pojavljivanja terma u određenom dokumentu i broj različitih dokumenata u kojima se pojavljuje određeni term.

34. Šta je ocena relevantnosti?

Ocena je mera koliko se dokument i upit poklapaju. Treba nam način za dodelu ocene svakom paru upit/do kument. Razmotrimo prvo upit sa jednim termom. Ako se term ne pojavljuje u dokumentu, ocena bi trebalo da bude 0. Što češće se term pojavljuje u dokumentu, ocena bi trebalo da bude veća.

35. Šta je frekvencija terma?

Frekvencija terma $tf_{t,d}$ terma t u dokumentu d definiše se kao broj pojavljivanja t u d tj. koliko puta se term pojavljuje.

	Ivanović D. [19]	Milosavljević B. [44]
digitalan	8	46
lucene	11	68
dokument	41	953
obrazovanje	0	0
pretraga	11	56

36. Šta je frekvencija dokumenta?

Koristimo frekvenciju dokumenta da uzmemo informativnost terma u obzir prilikom računanja ocene. df_t je oznaka za frekvenciju dokumenta i predstavlja broj dokumenata u kojima se pojavljuje term t.

term	df_t	idf_t
XMIRS	1	6
digitalizacija	100	4
nedelja	1000	3
analiza	10.000	2
ispod	100.000	1
i	1.000.000	0

37. Šta je tf-idf?

Jedna od najpoznatijih težina u oblasti pronalaženja informacija je tf-idf težina. tf-idf težina terma je proizvod njegove tf težine i njegove idf težine. Zavisna je od terma i od dokumenta, odnosno tf-idf težinu možemo dodeliti svakom termu t za svaki dokument d. Ova težina raste sa brojem pojavljivanja terma u dokumentu i sa retkošću terma u kolekciji. To je upravo ono što smo želeli da postignemo.

38. Objasniti kreiranje težinske matrice.

Pretpostavimo da imamo takvo pretprocesiranje teksta da su nam ostali samo sledeći termovi (ostali su izbačeni): indeksiranje, dokument, obrazovanje, multimedijalan. Brojačka matrica je data u nastavku. frekv. terma (broj). Kako ćemo kreirati težinsku matricu?

term	ID	MB	GS
indeksiranje	5	58	0
dokument	41	953	105
obrazovanje	0	0	1
multimedijalan	0	96	1

Prvo ćemo na osnovu datih $tf_{t,d}$ izračunati $w_{t,d}$ koristeći formulu.

$$w_{t,d} = \begin{cases} 1 + \log_{10} tf_{t,d} & \text{if } tf_{t,d} > 0 \\ 0 & \text{inače} \end{cases}$$

Nakon tog računa dobija se sledeća matrica: log frekv.

term	ID	MB	GS
indeksiranje	1,7	2,76	0
dokument	2,61	3,98	3,02
obrazovanje	0	0	1
multimedijalan	0	2,98	1

Izračunaćemo idf za svaki term koristeći formulu.

$$idf_t = \log_{10} \frac{N}{df_t}$$

term	idf
indeksiranje	0,176
dokument	0
obrazovanje	0,477
multimedijalan	0,176

N je 3 (imamo tri dokumenta u kolekciji), dft za term do kument je 3, za termove indeksiranje i multimedijalan je 2, a za term obrazovanje je 1. Dobijene vrednosti su sledeće: frekv. dok

Sada ćemo izračunati tf-idf za svaki term i dokument tako što ćemo pomnožiti prethodno izračunate tf i idf vrednosti prikazane u prethodne dve matrice. Dobijamo sledeću matricu: $tf_i df$

term	ID	MB	GS
indeksiranje	0,29	0,49	0
dokument	0	0	0
obrazovanje	0	0	0,48
multimedijalan	0	0,52	0,18

Sada je potrebno uraditi normalizaciju, odnosno podeliti svaki element matrice sa kvadratnim korenom zbira kvadrata svih elemeneta u toj koloni. Elemente u prvoj koloni ćemo deliti sa $\sqrt{0{,}29^2 + 0^2 + 0^2 +}$

$0^2 = 0{,}29$, elemente u drugoj ćemo deliti sa $\sqrt{0{,}49^2 + +0^2 + +0^2 + +0{,}52^2} = 0{,}71$, a elemente u trečoj koloni sa $\sqrt{+0^2 + +0^2 + 0{,}48^2 + 0{,}18^2} = 0{,}51$.

tf_idf & normalizacija

term	ID	MB	GS
indeksiranje	1	0,69	0
dokument	0	0	0
obrazovanje	0	0	0,94
multimedijalan	0	0,73	0,35

Na ovaj način smo dobili težinsku matricu.

39. Koje su razlike između Bulovog i Vektorskog modela pretraživanja?

Bulov Model:

Metod Pretraživanja: Koristi logičke operatore za striktno definisane upite.
Rezultati Pretrage: Binarni rezultati (relevantan ili nerelevantan).
Upotrebljivost: Efikasan za jednostavne i strogo definisane upite.
Fleksibilnost: Manje fleksibilan, teško se nosi sa nejasnim upitima.

Vektorski Model:

Metod Pretraživanja: Koristi težine termina i merenje sličnosti vektora za rangiranje relevantnosti.
Rezultati Pretrage: Rangirani rezultati po relevantnosti.
Upotrebljivost: Bolje performanse za složene i neprecizne upite.
Fleksibilnost: Veća fleksibilnost i bolje rangiranje rezultata.

40. Da li se relevantnost odgovora meri u odnosu na informacionu potrebu ili upit?

Zadovoljstvo korisnika se može meriti samo prema relevantnosti u odnosu na informacione potrebe, a ne upita, korisnik je zadovoljan samo ako je pronašao ono što je tražio.

41. Šta je preciznost (eng. precision)?

Preciznost P je udeo pronađenih relevantnih dokumenata u svim pronađenim dokumentima, odnosno u listi rezultata pretrage.

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

42. Šta je povrat (eng. recall)?

Povrat R je udeo pronađenih relevantnih dokumenata u svim relevantnim dokumentima koji postoje u kolekciji.

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

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

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

43. Šta je F mera i zašto je ona relevantnija od korišćenja preciznosti i povrata?

F mera omogućava da se meri kompromis između preciznosti i povrata.

$$\frac{2PR}{P + R}$$

Korišćenje samo preciznosti ili samo povrata može dovesti do jednostrane evaluacije sistema. Sistem može imati visoku preciznost, ali nizak povrat, što znači da mnogo relevantnih dokumenata nije pronađeno. Ili obrnuto, visok povrat, ali niska preciznost znači da je mnogo pronađenih dokumenata nerelevantno.

F mera pruža sveobuhvatniju sliku o performansama sistema, omogućavajući da se oceni kako sistem balansira između preciznosti i povrata.

44. Kako se može vršiti evaluacija performansi sistema za pretraživanje?

Evaluacija performansi može se koristiti za poređenje dva sistema, ali se često koriste i za unapređenje jednog sistema za pretraživanje. Za evaluaciju performansi sistema koriste se standardizovani test skupovi ili test skupovi specijalno kreirani za određeni sistem i za ove test skupove se računaju mere zasnovane na relevantnosti rezultata pretrage koje se koriste za ocenu kvaliteta sistema za pretraživanje - preciznost, povrat i F_1 . Povrat se teško meri kod sistema za pretragu velikih kolekcija kao što su veb pretraživači. Veb pretraživači obično koriste preciznost za najboljih k kao meru svojih performansi, na primer k = 10, ili koriste mere koje više vrednuju da je prvi pogodak bolji nego deseti. Takođe se koriste mere koje nisu zasnovane na relevantnosti: clickthrough za prvi pogodak, Laboratorijske studije ponašanja korisnika, A/B testiranje.

45. Šta je kapa mera?

Ocene relevantnosti su korisne samo ako su konzistentne među ocenjivačima. Konzistentnost među ocenjivačima možemo meriti kapa merom (k). Dakle, kapa je mera koliko se međusobno ocenjivači slažu i ova mera je dizajnirana za kategorične ocene.
P(A) = koji deo od ukupnog broja slučajeva se ocenjivači slažu, P(E) = koji deo slaganja bismo dobili slučajno.

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

46. Opisati A/B testiranje?

A/B testiranje ima za cilj testiranje jednog unapređenja sistema za pretraživanje. Uslov koji mora biti ispunjen da bi se moglo sprovoditi ovo testiranje je da postoji veliki pretraživač u pogonu sa velikim brojem korisnika i svakodnevnih upita.Većina korisnika pristupa staroj verziji sistema, a mali deo korisnika (recimo 1%) se preusmerava na novu verziju sistema koja ima unapređenja. Vrši se vrednovanje i stare i nove verzije sistema pomoću neke automatske mere, npr. clickthrough za prvi pogodak. Poređenjem ovih mera može se utvrditi da li unapređenje povećava zadovoljstvo korisnika. Jedna varijanta ove metodologije koja se ređe koristi je da se korisnicima ostavi mogućnost da sami izaberu staru ili novu verziju sistema. U ovom slučaju jako mali broj korisnika se odlučuje za novu verziju sistema.