Formiranje i pretraživanje tekstualnih baza podataka

Procenjuje se da u preduzećima 10% informacija čine strukturirani podaci koji se mogu efikasno čuvati u relacionim bazama podataka, doksu ostalih 90% različiti tekstualni dokumenti – priručnici, izveštaji, elektronska pošta, faksovi, WWW strane, prezentacije i slično. Od izuzetne važnosti je da informacioni sistemi omoguće efikasan pristup i ovim tipovima dokumenata.

Tekstualne baze podataka

Tekstualna baza podataka je kolekcija dokumenata za koju je obezbeđen efikasan metod pristupa i pretraživanja po sadržaju i po nekim drugim atributima dokumenata.

Arhitekture sistema za indeksiranje i pretraživanje tekstualnih informacija

Postoje dve osnovne arhitekture sistema za indeksiranje i pretraživanje tekstualnih informacija. To su:

1. proširenje relacionih baza podataka mogućnostima efikasnog skladištenja i pretraživanja velike količine tekstualnih podataka.
2. specijalizovani sistemi za indeksiranje i pretraživanje dokumenata.

Sistemi koji se baziraju na proširenju relacionih baza podataka

Sistemi koji se baziraju na proširenju relacionih baza podataka dokumente čuvaju ili u ćelijama tabela, ili na nekom drugom mestu u sistemu(na disku, na mreži) pri čemu se u ovom drugom slučaju , u ćelijama tabela čuvaju adrese dokumenata. Upitni jezik je SQL atipičan predsavnik sistema koji se baziraju na proširenju relacionih baza podataka je *Oracle.*

Specijalizovani sistemi za indeksiranje i pretraživanje dokumenata

Specijalizovani sistemi omogućavaju indeksiranje i pretraživanje dokumenata koji su smešteni u običnim datotekama fajl sistema. Ovo pretraživanje bazira se na upotrebi *Microsoft Index Servera* .

Microsoft Index Servera

Osnovne osobine *Index Servera* su:

- *Potpuna integrcija u Web server* – upiti se postavljaju iz standardnog WWW brouzera, gde se i prikazuju rezultati pretraživanja;
- *Indeksiranje po punom tekstu* – korisnik može dokumente da pretražuje porečima, frazama, pa čak i po kompletnim rečenicama;
- *Upiti po atributima dokumenata* – omogućeno je pretraživanje dokumenata po nekim njihovim atributima, kao što su ime autora, opis, veličina fajla i datum;
- *Neprecizni upiti* – korisnik može da koristi džoker znake i regularne izraze da bi pronašao sve gramatičke oblike reči;
- *Napredna pretraživanja* – upiti se mogu formirati kombinacijom primitivnijih uslova upotrebo operatora blizine (NEAR), numeričkih (<,=,>) i logičkih operatora (AND, OR, NOT);
- *Prilagodive forme za upite* – moguće je kreirati različite forme za postavljanje uslova i stranice rezultata;

- *Jednostavno održavanje – Index server* automatski ažurira bazu indeksa kada se dokumenti izmene, dodaju ili izbrišu;
- *Integrisana zaštita* – korisniku se može ograničiti pravo pristupa samo nekim dokumentima.

Indeksiranje dokumenata

Proces indeksiranja dokumenata sastoji se od sledećih faza:

- Filtriranje teksta
- Izdvajanje reči iz teksta
- Normalizacija
- Upis indeksa

Prvi korak u indeksiraju dokumenta jeste filriranje sadržaja. Iz dokumenta se izdvajaju delovi teksta koji se u daljem procesu indeksiranja mogu tretirati kao celina. Za svaki ip dokumenta oji se može indeksirati definiše se određeni filter.

Izdvajanje reči iz teksta obavlja komponenta koja se naziva *Word Breaker*. Ova komponenta vodi računa o specifičnostima kodnih rasporeda za pojedine jezike.

U fazi normalizacije teksta iz indeksa se isključuju reči koje se ne indeksiraju sva slova se pretvaraju u velika, što obezbeđuje da rezultat upita ne zavisi od tipa slova koja se koriste u upitu.

Pretraživanje

U najjednostavnijem slučaju za implementaciju aplikacije za pretraživanje dokumenata potrebno je kreirati tri fajla:

- HTML formu za postavljanje upita
- Definiciju upita
- Templejt za rezultate(HTX fajl koji je običan HTML fajl koji sadrži promenljive koje referišu podatke iz rezultata upita.

Sistem za pretraživanje informacija

Upotreba *Microsoft Index Servera* u kompleksnim informacionim sistemima nameće potrebu da se posebnim softverskim sistemom reše neki problemi. Ovi problemi se tiču:

- Način zadavanja složenih upita
- Održavanja korisničke sesije
- Distribucije upita većem broju servera
- Zaštite podataka
- Publikovanja dokumenata
- Integracije sistema u složene informacione sisteme.

Zadavanje složenih upita

Upiti se *Microsoft Index Serveru* zadaju preko komandnog jezika.Upit se u Sistemu za Pretraživanje Informacija (SPI) zadaje putem komandnog jezika, koji je u odnosu na Microsoft-ov jezik znatno uprošćen.

Korisnička sesija