RĪGAS TEHNISKĀ UNIVERSITĀTE
Datorzinātnes un informācijas tehnoloģijas fakultāte

INFORMĀCIJAS MEKLĒŠANAS SISTĒMAS INTERNETĀ

profesors
V. Zagurskis

SATURS

IEVADS

Pamata protokoli, kurus izmanto Internetā (turpmāk arī Tīklā), nav nodrošināti ar pietiekamām iebūvētam meklēšanas funkcijām, nerunājot nemaz par miljoniem serveru, kas atrodas tajā. HTTP protokols, ko lieto Internetā, ir labs tikai navigācijai, kur to izmanto tikai kā lappušu apskatīšanas līdzekli, bet ne pēdējo meklēšanai. Tas pats attiecās uz FTP protokolu, kas ir pat vēl primitīvāks nekā HTTP. Lielā informācijas pieauguma dēļ, navigācijas metodes ātri sasniedz savu funkcionālo iespēju robežas, nerunājot nemaz par efektivitātes robežām. Nenorādot konkrētus skaitļus, var teikt, ka nepieciešamo informāciju vairs nav iespējams iegūt uzreiz, jo tīklā pašlaik atrodas dokumentu miljardi un visi tie ir Interneta lietotāju rīcībā, turklāt to skaits pieaug eksponenciāli. Īsā laika sprīdī notika ļoti lielas informācijas izmaiņas. Pamata problēma ir, ka vienas, pilnas un funkcionālās šāda informācijas skaita atjaunošanas un ierakstīšanas sistēmas, vienlaikus pieejamas visiem Interneta lietotājiem visā pasaulē, nekad nav bijis. Tādēļ lai strukturētu informāciju, kas ir uzkrājusies Internetā, un apgādātu tās lietotājus ar ērtiem meklēšanas līdzekļiem, izveidoja meklēšanas sistēmas.

1. Meklēšanas sistēmas

Meklēšanas sistēmas parasti sastāv no trim komponentēm:

- aģents (zirneklis vai kroulers), kas pārvietojās pa Tīklu un ievāc informāciju;
- datu bāze, kas satur visu informāciju, ko savāca zirnekļi;
- meklēšanas mehānisms, ko cilvēki izmanto kā interfeisu bāzes lietošanā;

1.1. Kā strādā meklēšanas mehānismi

Meklēšanas un strukturēšanas līdzekļus, dažkārt sauktus par meklēšanas mehānismiem, izmanto, lai palīdzētu cilvēkiem atrast nepieciešamo informāciju. Aģentu, zirnekļu, krouleru un robotu tipa meklēšanas līdzekļus pielieto informācijas ievākšanai par dokumentiem, kas atrodas Internetā. Tās ir speciālās programmas, kuras nodarbojās ar lappušu meklēšanu Tīklā, izdala hiperteksta norādes uz lappusēm un automātiski indeksē atrasto informāciju datu bāzes veidošanai. Daži seko visām norādēm uz katras atrastās lappuses un pēc tam, savukārt, pēta katru norādi uz katras no jaunām lappusēm, utt. Daži ignorē norādes, kas ved pie grafiskiem, skaņas vai multiplikācijas failiem; citi ignorē norādes pie WAIS tipa datu bāzes resursiem; citi ir instruēti, ka jāizskata vispirms vispopulārākās lappuses.

- Aģenti – "visintelektuālākie" no meklēšanas līdzekļiem. Tie var vairāk nekā vienkārši meklēt: viņi var izpildīt pat transakcijas jūsu vārdā. Jau tagad aģenti var meklēt specifiskas tematikas saitus un atgriezt saitu sarakstus, izkārtotus pēc apmeklēšanas biežuma. Aģenti var apstrādāt dokumentu saturu, atrast un indeksēt citus resursu tipus, ne tikai lappuses. Tie var būt ieprogrammēti informācijas nolasīšanai no jau pastāvošām datu bāzēm. Neatkarīgi no informācijas, kuru aģenti indeksē, pārsūta to atpakaļ meklēšanas mehānisma datu bāzei.
- Informācijas vispārējo meklēšanu Tīklā realizē programmas, pazīstamas kā zirnekļi, Zirnekļi paziņo par atrastā dokumenta saturu, indeksē to un izdala gala informāciju. Vēl apskata virsrakstus, dažas norādes un atsūta indeksētu informāciju meklēšanas mehānisma datu bāzei.
- Krouleri izskata tikai virsrakstus un atgriež tikai pirmo norādi.
- Roboti var būt ieprogrammēti tā, lai pārietu pa dažādām norādēm dažādos dziļumos, izpildītu indeksāciju un pat pārbaudītu norādes dokumentā. To dabas dēļ tie var iestrēgt ciklos, tādēļ, izejot pa norādēm, tiem ir nepieciešami ievērojami Tīkla resursi. Tomēr ir

metodes, kuras ir paredzētas tam, lai neatļautu robotiem meklēšanu saitos, kuru īpašnieki nevēlās, lai tie būtu indeksēti.

Aģenti izdala un indeksē dažādus informācijas veidus. Daži, piemēram, indeksē katru vārdu atsevišķi visā sastopamā dokumentā, bet citi, savukārt, tikai 100 svarīgākos vārdus, dokumenta izmēru un vārdu skaitu tajā, nosaukumus, virsrakstus un apakšrakstus utt. Uzbūvētā indeksa veids nosaka, kādu meklēšanu var veikt ar meklēšanas mehānismu un interpretēt iegūto informāciju.

Aģenti var arī pārvietoties pa Internetu un atrast informāciju un vēlāk ievietot to meklēšanas mehānisma datu bāzē. Meklēšanas sistēmu administratori nosaka kādus un kāda tipa saitus aģentiem jāapmeklē un jāindeksē. Indeksēto informāciju nosūta meklēšanas mehānisma datu bāzei tāpat kā aprakstīts augstāk. Cilvēki var ievietot informāciju tieši indeksā, aizpildot īpašu formu tai nodaļai, kur viņi gribētu ievietot savu informāciju. Šos datus nosūta datu bāzei.

Ja kāds grib atrast informāciju, pieejamu Internetā, viņš apmeklē meklēšanas sistēmas lappusi un aizpilda formu, kurā ir aprakstīta viņam nepieciešamā informācija. Šeit var tikt izmantoti atslēga vārdi, datumi un citi kritēriji. Kritērijiem meklēšanas formā ir jāatbilst kritērijiem, ko izmanto aģenti indeksējot informāciju, kuru tie ir atraduši pārvietojoties pa Tīklu.

Datu bāze meklē uzdevuma priekšmetu, pamatojoties uz informāciju, kas ir norādīta aizpildītā formā, un izved atbilstošos dokumentus, kurus sagatavoja datu bāze. Lai noteiktu secību, kādā būs attēlots dokumentu saraksts, datu bāze pielieto aranžēšanas algoritmu. Ideālā gadījumā, dokumenti, kas visvairāk relevanti lietotāja pieprasījumam, būs sarakstā ievietoti pirmie. Dažādas meklēšanas sistēmas izmanto dažādus aranžēšanas mehānismus, tomēr relevances noteikšanas pamatprincipi ir sekojošie:

1. Pieprasījuma vārdu skaits dokumenta teksta saturā (t.i. html kodā).
2. Tegi, kuros šie vārdi izvietoti.
3. Meklēto vārdu izvietošana dokumentā.
4. Vārdu, pret kuriem noteikta relevance, pamatsvars kopējā vārdu skaitā dokumentā.
5. Šie principi pielietoti visās meklēšanas sistēmās. Bet zemāk parādītie pielietoti dažās, bet pietiekoši pazīstamās sistēmās (piem. AltaVista, HotBot).
6. Laiks – cik ilgi lappuse atrodas meklēšanas servera bāzē. Sākumā šķiet, ka tas ir diezgan bezjēdzīgs princips. Bet, ja iedomājas, cik daudz Internetā ir saitu, kas dzīvo maksimāli mēnesi! Ja saits eksistē pietiekoši ilgi, tas nozīmē, ka īpašnieks ir diezgan pieredzējis dotajā tēmā un lietotājam labāk derēs saits, kas jau pāris gadus vēsta pasaulei galda uzvedības noteikumus, nekā tas, kas parādījies pirms pāris nedēļām par šo pašu tēmu.

7. Citējamības indekss –norāžu daudzums no citām lappusēm, piereģistrētām meklēšanas bāzē, kas ved uz šo lappusi.

Datu bāze izvada šādā veidā aranžētu dokumentu sarakstu ar HTML un atgriež to cilvēkam, kurš izdarīja pieprasījumu. Dažādi meklēšanas mehānismi izvēlas arī dažādus iegūtā saraksta attēlošanas metodes – parāda tikai norādes; citi izvada norādes ar pirmiem teikumiem, ko satur dokuments vai dokumenta virsraksts kopā ar norādi.

Kad jūs noklikšķiniet uz viena jūs interesējoša dokumenta norādi, dokumentu pieprasa tam serverim, uz kura tas atrodas.

1.2. Meklēšanas sistēmu salīdzinošais apskats

Lycos. Lycos izmanto sekojošu indeksācijas mehānismu:

- Vārdiem virsrakstā <title> ir vislielākā prioritāte;
- Vārdi lappuses sākumā;
- Vārdi norādēs;
- Ja bāzes indeksā ir saiti, kur norāde norāda uz indeksējamo dokumentu, tad relevance pieaug.

Kā vairumā sistēmu, Lycos dod iespēju izmantot kā vienkāršo pieprasījumu, tā arī daudz izvērstāku meklēšanas metodi. Vienkāršajā pieprasījumā kā meklēšanas kritērijs tiek ievests teikums dabiskajā valodā, pēc kā Lycos izdara pieprasījuma normalizāciju, izdzēšot tā sauktos stop vārdus, un tikai pēc tam sāk tā izpildi. Gandrīz uzreiz tiek izdota informācija par dokumentu skaitu uz katru vārdu, bet vēlāk arī norāžu saraksts uz formāli relevantiem dokumentiem. Sarakstā katram dokumentam pretī tiek norādīta viņa tuvība pieprasījumam, vārdu skaits no pieprasījuma, kuri ir šajā dokumentā, un novērtējoša tuvība, kura var būt lielāka vai mazāka par formāli izskaitļoto. Pagaidām nevar ievadīt loģiskos operatorus rindā kopā ar terminiem, bet Lycos atļauj izmantot loģiku caur izvēlņu sistēmu. Tādu iespēju izmanto izvērstu meklēšanas formu būvēšanā, kas ir paredzēta lietotājiem, kuri jau ir iemācījušies strādāt ar šādu mehānismu. Tādejādi redzams, ka Lycos attiecās uz sistēmām ar pieprasījuma valodas tipu "Like this", bet ir paredzēta viņa paplašināšana arī uz citām meklēšanas organizācijas metodēm.

AltaVista. Šīs sistēmas indeksēšana tika veikta ar robota palīdzību. Pie tam robotam ir sekojošas prioritātes:

- Vārdiem tegā <title> ir visaugstākā prioritāte, atslēgu frāzes ir <Meta> tegos;
- Atslēgu frāzes, kuras ir lappuses sākumā;
- Atslēgu frāzes ALT – norādēs;

- Atslēgu frāzes pēc vārdu/frāžu skaita;

Ja tegu lappusē nav, tad izmanto pirmos 30 vārdus, tos indeksē un parāda apraksta vietā (tag description).

Visinteresantākā AltaVista iespēja ir izvērstā meklēšana. Šeit uzreiz jāpasaka, ka, atšķirībā no daudzām citām sistēmām, AltaVista atbalsta vienvietīgo operatoru NOT. Bez tam vēl ir operators NEAR, kas realizē konteksta meklēšanas iespēju, kad terminiem ir jāatrodas blakus dokumenta tekstā. AltaVista atļauj meklēšanu pēc atslēgu frāzēm, turklāt tai ir diezgan liela frazeoloģiskā vārdnīca. Bez visa pārējā, meklējot AltaVista, var uzdot laukuma vārdu, kurā ir jāsatiek vārds: hiperteksta norāde, applet, attēla nosaukums, virsraksts un daudz citu laukumu. Diemžēl, sīkāk aranžēšanas procedūra sistēmas dokumentācijā nav aprakstīta, bet ir redzams, ka aranžēšanu pielieto kā vienkāršā meklēšanā, tā arī pie izvērsta pieprasījuma. Reāli šo sistēmu var pieskaitīt pie sistēmas ar paplašinātu loģisko meklēšanu.

Yahoo. Dotā sistēma parādījās Tīklā viena no pirmajām, un šodien Yahoo sadarbojās ar daudziem informācijas meklēšanas līdzekļu izstrādātājiem, bet uz dažādiem tās serveriem izmanto dažādu programmatūru. Yahoo valoda ir pietiekami vienkārša: visus vārdus ir jāievada ar atstarpi, kur viņi tiek savienoti ar saikni AND vai OR. Uzrādot rezultātu, netiek uzrādīta dokumenta atbilstības pakāpe pieprasījumam, bet tiek pasvītroti vārdi no pieprasījuma, kuri tika sastapti dokumentā. Turklāt nenotiek leksikas normalizācija un nenotiek "kopējo" vārdu analīze. Labus meklēšanas rezultātus iegūst tikai tad, ja lietotājs zina, ka Yahoo datu bāzē šī informācija noteikti ir. Aranžēšanu veic pēc pieprasījuma terminu skaita dokumentā. Yahoo pieder pie parasto tradicionālo sistēmu klases ar ierobežotām meklēšanas iespējām.

OpenText. Informācijas sistēma OpenText pārstāv plašu komercializētu informācijas produktu Tīklā. Tās apraksti vairāk līdzinās reklāmai, nekā darba informatīvajiem norādījumiem. Sistēma atļauj veidot meklēšanu, izmantojot loģiskos konektorus, toties pieprasījuma izmērs ir ierobežots ar trim terminiem vai frāzēm. Dotajā gadījumā runa iet par paplašinātu meklēšanu. Uzrādot rezultātus, tiek paziņots par dokumenta atbilstību pieprasījumam un tā izmērs. Sistēma atļauj arī uzlabot meklēšanas rezultātus tradicionālās loģiskās meklēšanas stilā. OpenText varētu pieskaitīt pie tradicionālām informacionālam meklēšanas sistēmām, ja ne aranžēšanas mehānisms.

InfoSeek**.** Šajā sistēmā indeksu veido robots, kas indeksē ne visu saitu, bet tikai norādīto lappusi. Turklāt robotam ir sekojošas prioritātes:

- Vārdi virsrakstos <title> ir vislielākā prioritāte;
- Vārdi tegos keywords, description un atkārtošanās/esamības biežums pašā tekstā;
- Atkārtojoties vienādiem vārdiem blakus, tos izmet no indeksa;
- Atļauj līdz 1024 simbolus tegam keywords, 200 simbolus tegam description;

- Ja tegi netika izmantoti, indeksē pirmos 200 vārdus lappusē un izmanto tos kā aprakstu;

InfoSeek sistēma ir apgādāta ar diezgan attīstītu informācijas meklēšanas valodu, kas ļauj ne tikai norādīt kādi termini jāsatiek dokumentā, bet arī uzsvērt tos. Sasniedz to ar speciālo zīmju palīdzību: "+" – terminam obligāti jābūt dokumentā, " - " – termins nedrīkst būt dokumentā. Bez tam, InfoSeek atļauj izpildīt tā saucamo konteksta meklēšanu. Tas nozīmē, ka izmantojot speciālo pieprasījuma formu, var pieprasīt secīgu kopējo vārdu sastopamību. Tāpat var norādīt, ka dažiem vārdiem ir jāsatiekas ne tikai vienā dokumentā , bet pat vienā paragrāfā vai virsrakstā. Ir atslēgas frāžu, kas izskatās kā viens vesels, norādes iespējas. Aranžēšana pie izvades tiek realizēta pēc pieprasījuma terminu skaita dokumentā, pēc pieprasījuma frāžu skaita, neskaitot kopējos vārdus. Visus šos faktorus izmanto kā ieliktās procedūras. Rezumējot jāsaka, ka InfoSeek pieder pie tradicionālām sistēmām ar meklēšanas elementu uzsvēršanu .WAIS**.** Wais pārstāv vienu no izvērstākām meklēšanas sistēmām Internetā, kurā nav realizētai tikai meklēšana pēc neskaidriem lielumiem un varbūtēja meklēšana. Atšķirībā no daudzām meklēšanas mašīnām, sistēma ļauj noteikt ne tikai ieliktos loģiskos pieprasījumus, nolasīt formālo relevanci pēc dažādām tuvības pakāpēm, uzsvērt pieprasījuma terminus un dokumentus, bet arī realizēt pieprasījuma korekciju pēc relevances. Sistēma arī ļauj izmantot terminu nogriešanu, dokumentu sadali laukumos un sadalīto indeksu vešanu. Tādēļ tieši šī sistēma izvēlēta par pamata meklēšanas mašīnu enciklopēdijas "Britannica" realizācijai Internetā.

2. PROFESIONĀLA MEKLĒŠANA INTERNETĀ: PILNĪGUMS, DROŠUMS, ĀTRUMS

2.1. Profesionālās meklēšanas raksturs.

Tātad, atšķirībā no situācijas, kad jūs meklējat kaut ko sev, profesionālā meklēšana ir paredzēta kāda pasūtījuma izpildei, ar izrietošo no tā atbildību pasūtītāja priekšā. Šī atbildība ir arī avots trim pamata prasībām:

- Resursu pilnās ietveršanas kontrole;
- No tīkla iegūtās informācijas ticamības kontrole;
- Liels meklēšanas ātrums.

Tātad, ja jūs esat pasūtītāja lomā, tad jums ir tiesības pieprasīt no meklētāja, izņemot rezultātus, vēl arī garantiju par iepriekš nosauktiem punktiem. Tādas garantijas, protams, var dot tikai cilvēks, kurš ir pietiekami informēts par informācijas plūsmas sadales un kustības sīkumiem Internetā.

2.1.1. Resursu aptveršanas pilnības kontrole

Resursu aptveršanas pilnības kontrole ir pilntiesīga prasība, ja jūs risināt uzdevumu, kas ir pretējs uzdevumam, kurš skan kā "atrast vismaz kaut ko".

Pilna mēroga informācijas meklēšana kādā jautājumā no Interneta daudzos gadījumos izved meklētāju aiz plaši apgūta Web – plašuma robežām, novedot pie telnet pieejamiem datu bāzēm, reģionālām telekonferencēm un citām informācijas glabātuvēm. Visu šodien eksistējošo tipu pamatresursu zināšana, to informatīvās piepildīšanās tehniskās un tematiskās specifikas un pieejas īpašību saprašana, kļūst par nepieciešamu nosacījumu veiksmīgai meklēšanas darbu plānošanai un izpildei.

2.1.2. Informācijas ticamības kontrole

Informācijas, iegūtas no Tīkla, meklēšanas rezultātā, ticamības kontrole, var realizēt ar dažādiem līdzekļiem. Īsi aprakstīsim iespējās, kuras piedāvā pats Tīkls. Tā, par tradicionālām pārbaudes metodēm var uzskatīt informācijas avotu, kas ir alternatīvs dotajam, lokalizāciju, faktiskā materiāla pārbaude, tā izmantošanas ar citiem avotiem, biežuma noskaidrošana, dokumenta statusa un mezgla, uz kura to atrod meklēšanas sistēmas, reitinga noskaidrošana, informācijas iegūšana par materiāla autora kompetenci un statusu ar speciālo meklēšanas servisu palīdzību, mezgla organizācijas atsevišķo elementu analīze ar mērķi novērtēt tā apkalpošanas speciālistu kvalifikāciju un citi.

2.1.3. Meklēšanas realizēšanas ātrums Tīklā

Meklēšanas realizēšanas ātrums Tīklā, ja neņem vērā lietotāja pieslēguma tehniskos raksturojumus, atkarīgs no meklēšanas procedūras pareizas plānošanas un pieredzes darbā ar izvēlēto resursu. Ar meklēšanas darbu plāna sastādīšanu saprot meklēšanas servisu un instrumentu, kas atbilst uzdevuma specifikai, izvēli, un, kas ļoti svarīgi, to pielietošanas secības, atkarībā no gaidāmā rezultāta. Pēc pieejas iegūšanas pie atbilstošā resursa, pirmā vietā ir māka ātri izprast tā struktūru un navigācijas metodes. Izpildāmo darbību motorika, prasmīga meklēšanas līdzekļu informācijas apstrādes iespēju savietošana lokālai klienta programmai un serverim, ir meklētāja nepieciešamās iemaņas.

2.2. Resursu tipi Internetā

Šodien informācija Internetā ir pieejama no dažādiem avotu tipiem. Plānot meklēšanu bez saprašanas par to spektru un funkcionēšanas īpašībām nav iespējams. Resursu pamata tipu saraksts ir dots 1.1.attēlā. Faktiski jautājums ir stādīts plašāk – par informācijas attēlošanas, pārsūtīšanas un apstrādes pamata metodēm Tīklā.

Interneta informācijas un komunikācijas pamata resursi
• Elektroniskais pasts un pasta roboti; • Telekonferenciju Usenet globālā sistēma, reģionālās un specializētas telekonferencijas; • Nosūtīšanas saraksti; • Lietotāju komunikācijas onlaina (tiešsaistes) līdzekļi; • Cilvēku un organizāciju meklēšanas līdzekļi; • Hytelnet datu bāzes; • Failu arhīvu FTP sistēma, globālās un reģionālās ietveršanas meklēšanas sistēmas FTP arhīvos; • Gopher datu bāzes un meklēšanas sistēma Veronika; • Hiperteksta informācijas sistēma World Wide Web (WWW); • Resursu katalogi – globālie, lokālie, specializētie (WWW vidē); • Meklēšanas mašīnas, vai automātiskie indeksi – globālie, lokālie, specializētie (WWW vidē); • Banneru sistēmas (WWW vidē); • Aktīvie informācijas kanāli (vidē);

1.1.att. Interneta informācijas un komunikācijas pamata resursi

Norādītā tipa resursu pieejas īpašības aprakstītas daudzās rokasgrāmatās. Īsi aprakstīsim katru resursa tipu, pievēršot uzmanību tam, kādu slodzi nes resurss veicot meklēšanu Tīklā.

Elektroniskais pasts un pasta roboti. Atsevišķas personas vai organizācijas elektroniskā pasta adresi tradicionāli izmanto īpašnieka identifikācijai. Tīkla komunikācijas resursos, lietotāju onlaina komunikācijas līdzekļos un telekonferenču sistēmās tas ir nepieciešams atribūts katram dalībniekam. Speciālā URL shēma mailto atļauj ielikt Web lappusē hipernorādi uz e-mail, kas automātiski atver pasta klientu. Tādā veidā tā tiek plaši pielietota Tīklā. Pašas adreses brīvi indeksē ar meklēšanas sistēmām un ir pieejamas meklēšanai caur kopīgas nozīmes

meklēšanas mašīnām. AltaVista, piemēram, rāda, ka elektroniskā pasta adreses ir sastopamas gandrīz 100 miljonos no150 miljoniem Web lappušu indeksētiem dokumentiem.

E-mail adreses aktīvi uzkrāj speciālās cilvēku un organizāciju meklēšanas sistēmās. Lieku neērtību, pie meklēšanas pēc e-mail, sagādā tas, ka pie adreses saņemšanas atļauj lietotāja reģistrāciju zem pseidonīma. Šī prakse ir bieži pielietota uz serveriem, kas piedāvā bezmaksas pastkastītes.

Pasta roboti – tās ir speciālās programmas, kas ir spējīgas atbildēt ar noteiktām darbībām uz komandām, kas tiem tiek nosūtītas pa elektronisko pastu. To pamatuzdevums – datu nosūtīšana pēc pieprasījuma gadījumā, ja tie nav pieejami ar citu metodi, kā arī darbības alternatīvā režīmā on-line ar kādu no pazīstamiem resursiem. Pie meklēšanas pasta robotus bieži izmanto kā starpniekus informācijas saņemšanai. Dažreiz jāsastopas ar to, ka tas ir vienīgais veids, kā var saņemt nepieciešamos datus.

Telekonferenču Usenet globālā sistēma, reģionālas un specializētas telekonferences. Sistēma uzbūvēta pēc elektronisko ziņojumu dēļu principa, kur lietotājs var izvietot savu informāciju vienā no ziņu tematiskajām grupām. Pēc tam šo informācija nodod lietotājiem, kuri ir pierakstījušies uz šo grupu. Usenet ziņu grupu kopējais skaits pārsniedz 20 tūkstošus un ziņas par tiem var atrast, piemēram uz Yahoo. Visi viņi vienlaicīgi netiek atbalstīti ne ar vienu serveri, tādēļ grūtāk ir atrast nevis atbilstošās grupas nosaukumu, bet gan telekonferences serveri, no kura to varētu ielādēt. Usenet ir atslēgvārds tieši globālai telekonferences sistēmai. Reģionālās un specializētas sistēmas arī ir izplatītas. Resursam ir vislielākā nozīme informācijas ātrai uzkrāšanai kādā šaurā jautājumā, bet pie meklēšanas, biežāk privātas, neoficiālās informācijas iegūšanai. Lūk, kāds piemērs. Viens no referentiem saņēma uzdevumu nodrošināt vienas kompānijas delegācijas uzturēšanās Londonā "tehnisko" daļu. Šai gadījumā nepieciešamais standarta ziņu kopums – transports, viesnīca, laika prognoze, pēdējās pilsētas ziņas, kā arī komandējuma dalībnieku personiskās vēlmes. Informācijas lielākā daļa tika paņemta no Web mezgliem, lokalizētiem ar Yahoo un AltaVista meklēšanas sistēmu palīdzību. Tomēr uz dažiem jautājumiem, tādiem kā automašīnas noma un Londonas sabiedriskā transporta maršruti, atbildes Web telpā neeksistēja. Ar servera Deja News (http://wmod.dejanews.com) palīdzību, kas ir telekonferences sistēmas Web vārteja, referents atrada divas britu reģionālās ziņu grupas – uk.transport.london un uk.local.london. Pēc sava lūguma izskaidrošanas, visa nepieciešamā informācija tika iegūta vienas dienas laikā.

Nosūtīšanas saraksti. Tie paredzēti vairāk vai mazāk sistemātiskai informācijas izsūtīšanai pa elektronisko pastu. Ja lietotājs var pats ievietot informāciju nosūtīšanas sarakstā, tad tas sāk atgādināt telekonferences sistēmu, bet nepieprasa speciālo klientu. Nelielo pēc adrešu aptveres un šauri specializēto vai reklāmas nosūtīšanas sarakstu Tīklā ir milzīgs daudzums. Ja

nerunājam par kaut kādām speciālām interesēm, tad minētā informācija ir nepieciešama meklētājam, galvenokārt, lai būtu kursā par pēdējiem notikumiem Internetā. Valdīšana par tīkla leksiku plašā tēmu spektrā un informētība par lielākiem projektiem, kas tiek realizēti Tiklā, kas atrodami nosūtīšanas sarakstos, ļauj daudz rezultatīvāk uzstādīt meklēšanas pieprasījumus.

Onlaina lietotāju komunikācijas līdzekļi (chat, ICQ, un citi) paredzēti informācijas apmaiņai starp diviem un vairāk Tīkla lietotājiem reāla laika režīmā, izmantojot speciālo čata serveru. Par šādas apmaiņas daļu var kļūt teksta dialogs, grafikas pārsūtīšana tās veidošanas procesā, balss vai video sakari, failu apmaiņa. Ilgu laiku šī tipa resursi reti tika izmantoti meklēšanas uzdevumu risināšanai, tomēr situāciju mainīja 1996.gadā ievestais jaunais serviss ICQ. Atšķirībā no agrāk pastāvošiem čatiem, kur dalībnieku reģistrācijai bija anonīms raksturs un tā darbojās tikai sakaru seansa laikā, ICQ izstrādātāji piedāvāja katram lietotājam reģistrācijas numuru – identifikatoru, kurš saglabātos pie viņa pastāvīgi. Risinājumam bija grandiozās sekas cilvēku sazināšanās jomā ar datoru. Unikālais ICQ numurs var parādīties uz vizītkartēm blakus telefona numuram, elektroniska pasta un mājas lappuses adresei. Pie cilvēku un organizācijas meklēšanas var sekmīgi izmantot ICQ meklēšanas dienestu, kas kļūst pieejams uzreiz pēc ICQ klienta numura uzstādīšanas datorā.

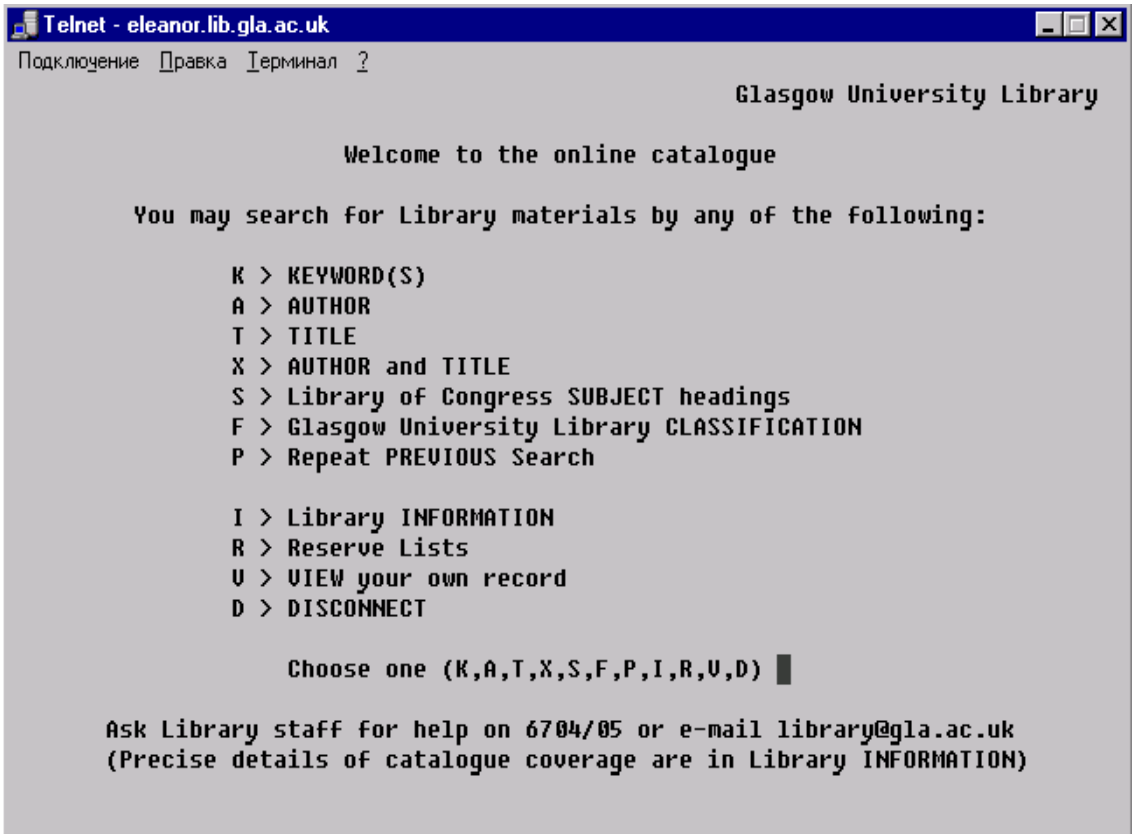
Vēl daži vārdi par čata serveriem. Parasti neliels serveru saraksts ir jau iešūts izmantojamā klienta programmā, kā, piemēram, programmā Microsoft NetMeeting.

Čatu reģistrācijas sarakstos parasti ir ziņas par dalībnieku dzīvesvietu, un pēdējā bieži vien ir norādīta nepareizi. Čata resursi, pat savā anonīmajā variantā, piesaistījuši uzmanību ar to, ka tie ļauj iegūt informāciju no pirmām rokām no planētas konkrētas valsts, reģiona un pilsētas pārstāvja.

Cilvēku un organizāciju meklēšanas sistēmas mūsdienu Tīklā raksturo ar diviem svarīgiem momentiem: vairākums no šiem resursiem jau ir pārnests uz Web serveriem un aizvien lielāku klātbūtni tajos iegūst informācija par cilvēkiem un organizācijām, kuriem nav tieša vai vispār nekāda sakara ar Internetu. Sakarā ar pēdējo apgalvojumu ir zināmi fakti, kad Tīklā parādījās kā atsevišķu organizāciju tā arī veselu reģionu telefonu, adrešu un citas datu bāzes. Tomēr tāds lietotāja tīkla identifikators, kā e-mail adrese, paliek par dominējošo meklēšanas atribūtu daudziem šī tipa servisiem. Par viņu datu bāzes papildināšanas avotiem kļūst telekonferenču materiāli, Web serveri, kā arī lietotāju patstāvīgā reģistrācija. Pie tiem pievieno sistēmas, kas specializējās uz meklēšanu. Piemēram, pēc ICQ numura vai lietotāju mājas lappusēm (Ahoy! dienests). Kopā ar pārorientētiem un WWW servisiem, Tīklā turpina darbu viens no vecākiem šāda tipa meklēšanas dienestiem Whois, kas ir pieejas pa protokolu telnet no servera whois.internic.net pēc ieiešanas ar login: whois.

Bieži rodas mēģinājumi noskaidrot šī uzdevuma meklēšanas servisu reitingu. Tā, pēc žurnāla PC Magazine (http://zdnet.com/pcmag) izpētīšanas, vislielākā popularitāte Tīklā starp Eiropas un Ziemeļu Amerikas lietotājiem ir elektronisko pastu adrešu meklēšanas dienestam Four11 (http://www.four11.com) , kas ir izvietota Yahoo portālā. Tomēr prakse liecina, ka meklēšanas uzsākšana tieši no tā negarantē veiksmi. Visiem šiem dienestiem ir viens nopietns trūkums – tie nepārstāv kopīgu, administrēto sistēmu, bet ir tikai haotisks, pēc sveša vērotāja uzskata, piepildāmo informācijas mezglu krājums. Kā sekas tam ir tas, ka pareizi saplānot meklēšanas procedūru un izvietot prioritātes atsevišķas personas meklēšanā kļūst ļoti sarežģīti. Dažos gadījumos daudz efektīvāk cilvēku meklēt pēc viņa pēdām Tīklā – publikācijām, dienesta vietas utl, ar kopējās nozīmes meklēšanas sistēmu izmantošanu.

Hytelnet datu bāzes, kas ir pieejamas pa telnet protokolu, daudzos gadījumos attēlo absolūti unikālu informāciju, galvenokārt pēc Eiropas un Amerikas universitāšu bibliotēku, kā arī valsts iestāžu, katalogiem. Visiespaidīgāko šā tipa datu bāžu sarakstu, kas ietver 1600 vienības, var atrast uz Web servera pēc adreses http://www/lights.com/hytelnet/. Katrai no tām ir sava oriģināla navigācijas un meklēšanas sistēma, kas tiek realizēta caur komandām, kuras tiek ievadītas no klaviatūras burtciparu režīmā. Līdzīga interfeisa piemērs, ar kuru nav pazīstama lielākā šodienas lietotāju daļa, ir dots 1.2.attēlā.



1.2.att. Interfeisa piemērs, kas ir pieejams pa telnet protokolu, bibliotēkas datu bāzei Glasgow University (UK)

Failu arhīvu FTP sistēma, globālās un reģionālās ietveršanas meklēšanas sistēmas FTP arhīvos. Šī tipa resursi nepiekāpās bez ierunām zem Web tehnoloģiju spiediena, kā vairākums pārējo. Viens no iemesliem – lielais informācijas apjoms, kas gadu desmitiem uzkrājies FTP arhīvos datoru sistēmu ekspluatācijas laikā, joprojām ir vērtīgs speciālistiem. Sociālā pieprasījuma pēc tā pārnešanas pilnā mērogā Web telpā nepastāv. Cits iemesls slēpjas pieejas, navigācijas un failu pārsūtīšanas vienkāršībā pa FTP. Tādejādi vai savādāk šodien FTP resursi ir pieprasīti un pat attīstās ne tikai savā vienīgā globālā meklēšanas sistēmā Archie (viena no stabilākām pieejas vārtejām ir http://ftpsearch.ntnu.no/), bet arī reģionālās sistēmās. FTP arhīvi – tas pirmkārt ir programmatūras nodrošinājuma avots, kas veiksmīgi konkurē ar Web mezgliem, specializētiem programmu kolekciju reprezentēšanai un pārdošanai. Atšķirībā no Web mezgliem, tajos biežāk var sastapties ar autortiesību pārkāpšanu programmu pirātkopiju un daļēju materiālu veidā, ko citos mezglos pārdod par naudu. Kā FTP servisa ēnas pusi var minēt bīstamību inficēties ar vīrusu no nepārbaudīta avota. Kādas informācijas meklēšana jāsāk no FTP meklēšanas sistēmas? Universālā atbilde ir vienkārša: par cik kā atslēgvārds formējot pieprasījumu ir teksts, kas ietilpst faila vai kataloga nosaukumā uz FTP servera, tad labākus rezultātus var iegūt meklējot informāciju, kura ir noformēta faila veidā, kuram vai nu jau ir kāds noteikts nosaukums, vai pastāv reāla iespēja to uzminēt. Lietišķā FTP meklēšanas pielietojuma piemēru ir diezgan maz. Meklētājs, kas meklēja vienu no amerikāņu standartiem ASTM materialzinībās ar HotBot palīdzību, ātri lokalizēja galveno Web serveri. Tur viņam izdevās noskaidrot standarta precīzo nosaukumu. Standarta pilns apraksts tika piedāvāts par maksu, bet īsa anotācija – bezmaksas. Tehnisko iemeslu dēl anotācija uz servera nebija pieejama. Cilvēks pieņēma lēmumu izpētīt FTP arhīvus ar meklēšanas sistēmas palīdzību un izmantot alfabēta – ciparu secību, kas kodēja materiāla nosaukumu. Drīz vien tika atrasta standarta versija, tuva pilnai, kas risināja problēmu. Informācijas patiesums meklētajam izraisīja šaubas, bet to vienkārši varēja noskaidrot speciālisti.

Gopher datu bāzes un meklēšanas sistēma Veronika, kas skanē Gopher telpas resursus, pašlaik pārstāj spēlēt nozīmīgu lomu Interneta informācijas laukā. Tomēr visas pasaules Gopheru māte – serveris uz kura ir reģistrēti vairāki Tīkla Gopher serveri (gopher://gopher2.tc.umn.edu), paliek darba stāvoklī pat šodien. Iziet uz tā vai cita Gopher servera gadās caur norāžu kolekciju uz Web lappusēm, un caur "papīra" Dzeltenām lapām. Parasti, ja Gopher serveris vēl strādā, tad uz tā failiem ir norāde ar Web mezgla adresi, uz kuru pārnesta visa informācija.

Hiperteksta informācijas sistēma World Wide Web (WWW) un tās tehnoloģijas mūsu dienās ir visnozīmīgākās Tīklā un turpina celties. Pēc savas navigācijas bildes WWW faktiski nokopēja Gopher resursus, bet vienas sīkas detaļas sekas maz kas varēja paredzēt. Šī detaļa, Web lappuses izmantošana kā viegli veidojams objekts, kurā tiek iemontēti citi vienkāršāki objekti, kas paredzēti vienlaicīgai attēlošanai. Tas, ka šodien pēdējo sarakstā ir teksts, hipernorādes, grafika, multimedija, programmas kods, dialoga formas un daudz kas cits, gala rezultātā arī noteica WWW plašo komerciālo pielietojumu. Tīkls piespieda Web telpas meklēšanas sistēmas pakļaut sev un faktiski noteica to attīstības tendences. No vienas puses runa ir par to, ka pie resursu indeksēšanas, meklēšanas sistēmas detalizētāk apstrādā Web lappušu laukumus, kurus formē HTML valodas konteineri. No otras puses intensīvi attīstīti tiek tie informācijas meklēšanas elementi, kas atbalsta meklēšanu šo laukumu iekšpusē. Šodien var konstatēt dziļu meklēšanas sistēmu un WWW resursu integrācijas pakāpi uz vienas tehnoloģijas bāzes. Bez tam WWW bāzes milzīgais izmērs asi nostāda jautājumu par veselu virkni identisku paralēlo meklēšanas sistēmu eksistenci, kas apkalpotu lietotāju intereses.

Resursu katalogi – globālie, reģionālie un specializētie (WWW vidē), pārstāv Tīklā izvietotas datu bāzes ar resursu adresēm un ar dažādu uzkrātas informācijas mērogu un tematiku. Parasti tiem ir hierarhiskā struktūra, pārvietojoties kurā var lokalizēt nepieciešamo objektu. Informācijas uzkrāšanas ātrums šādās sistēmās izradās salīdzinoši mazs, par cik resursu klasifikācijā ir paredzēta cilvēka līdzdalība. Meklētājam informācijas saņemšana no pazīstama kataloga, vienmēr ir patiesuma garantija. Veicot vairāk vai mazāk standarta meklēšanas uzdevumu, tieši katalogs, nevis meklēšanas mašīna, kļūst par starta laukumu meklēšanas sākumam.

Meklēšanas mašīnas, vai automātiskie indeksi – globālie, lokālie, specializētie (WWW vidē) pārstāv jaudīgas informācijas meklēšanas sistēmas, kas tiek izvietotas uz brīvpieejas serveriem. To speciālās programmas – roboti, jeb zirnekļi, automātiskā režīmā nepārtraukti skanēs pēc uzdotiem algoritmiem Tīkla informāciju, veicot dokumentu indeksāciju. Turpmāk pēc izveidotām indeksa datu bāzēm meklēšanas mašīnas piedāvā lietotājam pieeju pie uz Tīkla mezgliem sadalītas informācijas. Tas tiek realizēts, izdarot meklēšanas pieprasījumus atbilstoša interfeisa ietvaros. Pēdējās meklēšanas mašīnu iespēju izpētes rāda, ka pat tik jaudīgo mašīnu, kā AltaVista un HotBot, Pasaules Tīkla resursu aptveres pilnība nepārsniedz 30%. Meklēšanas procedūras plānošana WWW vidē ir netriviāla, un to jāapskata atsevišķi.

Banneru (reklāmkarogu) sistēmas (WWW vidē) paredz speciālo objektu – banneru izvietošanas variantus. Banners parasti ir neliels grafiskais attēls, kas ar reklāmas mērķi tiek izvietots uz Web mezgla, kas pieņem reklāmu. Banneri aizsūta lietotāju pēc hipernorādes uz

reklāmas devēja serveri un bieži vien tiem nav nekāda sakara ar lappuses pamata saturu. Banneri netiek tieši lietoti pie meklēšanas, bet tie ir labi Tīkla informācijas tirgus stāvokļa indikatori.

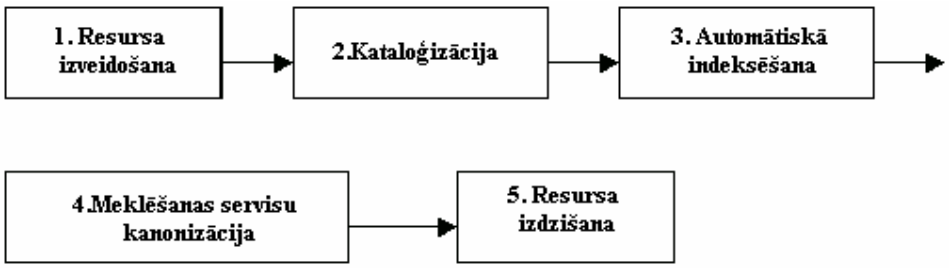
Aktīvie informācijas kanāli (WWW vidē) pārstāv specializētus Web serverus, kas ir paredzēti datu nokļūšanai tieši lietotāja darba vietā. Šī tipa resursus pieņemts saistīt ar push tehnoloģiju (informācijas izbīdīšanas tehnoloģija). Faktiski aktīvais Web kanāls ir periodiski atjaunojamo datu informācijas avots. Ir iespējams kā parakstīties uz kanālu, tā arī pārtraukt parakstīšanos, kas daudziem atgādina darbu ar nosūtīšanas sarakstiem. Kanālu atbalsta metodika šodienas pamata brauzeriem (pārlūkiem) Netscape Communicator un Internet Explorer ir dažāda. Ar kanālu informāciju, pēc tās atjaunošanas, ir iespējams vēlāk iepazīties autonomajā režīmā. Pati tehnoloģija neieguva gaidīto plašo izplatību un meklēšanas problēmas kontekstā nespēlē ievērojamu lomu.

2.3. Interneta resursi caur meklēšanas serveru prizmu

Starp Interneta lietotājiem viegli norādīt divas kategorijas. No vienas puses tie ir resursu izstrādātāji visplašākajā šī vārda nozīmē, no tehniskā personāla līdz žurnālistiem t.i. autoriem, kuri nogādā informāciju Tīklā. No otras puses - aktīvie informācijas plūsmas lietotāji. Informācijas meklēšanas darbība kļūst par lietotāju sfēras sastāvdaļu.

Izstrādātāju centieni saprast lietotāju intereses izskatās vairāk nekā dabiski. Tomēr meklēšanas uzdevumu risināšanas efektīvie piegājieni slēpjas tieši pretējā – meklēšanas interešu, nodomu un tehnisko risinājumu detalizētākā izprašanā, ko kultivē izstrādātāji. Tādēļ, izskatot Tīkla resursu pamattipus, minēti arī tie, kas pagaidām lielākā mērā ir pievilcīgi informācijas piegādātajiem. Dažu resursu loma meklēšanas uzdevumos sākumā liekās nesvarīga, bet šis stāvoklis var mainīties.

Internet – tehnoloģiju attīstības vēsture rāda, ka meklēšanas servisu, kas apkalpo noteikta tipa informācijas resursu, stāvoklis ir tieši saistīts ar tā dzīves cikla fāzi, kas ir parādīts 1.3.att.



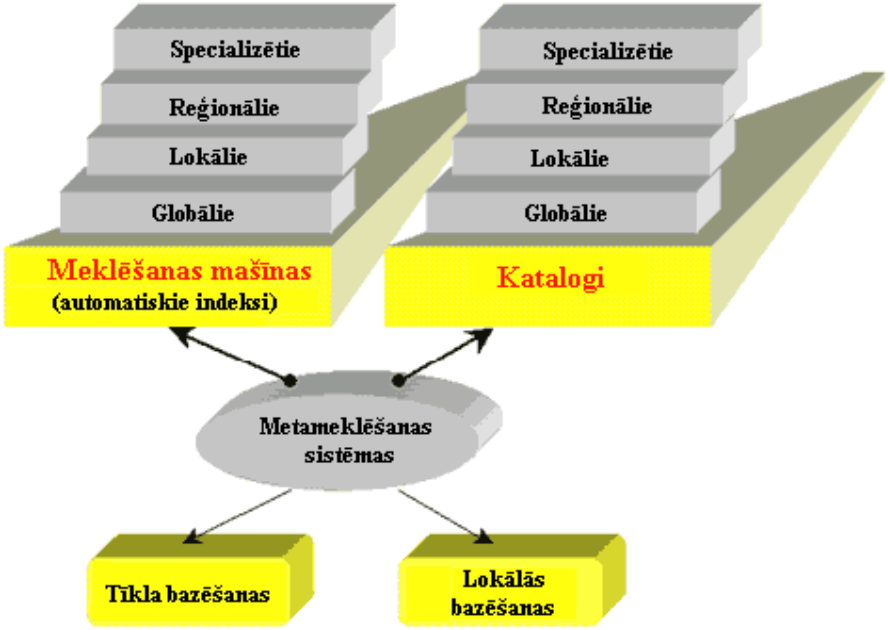
1.3.att.Tīkla informācijas resursa dzīves cikla saistība ar meklēšanas servisu attīstības dinamiku.

Īsi paskaidrosim dzīves cikla pamata elementus. Kataloģizācija, kā norāžu kolekcijas noformēšana un palielināšana uz dota tipa resursiem, seko uzreiz pēc resursa uzstādīšanas. Automātiskās indeksēšanas serviss parasti sāk formēties, kad resursa informācijas masa sasniedz kādu kritisko vērtību. Tālāk notiek identisko meklēšanas servisu – katalogu un indeksu, kas apkalpo resursu, konkurences fāze. Kanonizācija praktiski apstādina šo procesu, nododot prioritāti vienai vai vairākām meklēšanas sistēmām. Noslēdzošā stadija – resursa izdzišana – tiek raksturota ar informācijas masas pārvietošanu uz citu resursu tipa funkcionēšanas laukumu līdz resursa pilnīgai izzušanai.

Mēģināsim izskatīt pēc 3.att. shēmas tādas informācijas sistēmas, kā Telnet, FTP, Gopher, un WWW. Tā ir acīmredzams, ka WWW resursi pašlaik pārdzīvo dzīves cikla smaili starp cikla 3 un 4 fāzi. FTP arhīvi šobrīd atrodas kanonizācijas fāzē. Gopher un Informācijas meklēšanas darbi resursā, kas pārdzīvo automātisko indeksu strauju attīstīšanos, var būt vienlaicīgi daudzsološi un problemātiski. Telnet datu bāzes raksturo ar izdzišanas fāzi. Tomēr, lai kādu fāzi nepārdzīvotu resurss, tas vienmēr var saturēt unikālu informāciju un tāpēc, organizējot meklēšanu Tīklā, ar to jārīkojas saudzīgi.

3. PROFESIONĀLĀ MEKLĒŠANA INTERNETĀ: MEKLĒŠANAS PROCEDŪRAS PLĀNOŠANA

3.1. Interneta meklēšanas servisu struktūra. Meklēšanas mašīnas un katalogi



2.1.att. Interneta meklēšanas servisu organizācija

Pēc shēmas 2.1.att. reālos informācijas nesējus Internetā pārstāv meklēšanas mašīnas (automātiskie indeksi) un katalogi. Tā kā tie, kaut arī ar dažādiem paņēmieniem, patstāvīgi nodrošina visus informācijas apstrādes etapus, no tās saņemšanas mezgliem – pirmavotiem līdz meklēšanas iespēju piedāvāšanai lietotājam, tos bieži sauc par autonomām sistēmām.

Autonomās meklēšanas sistēmas var atšķirties pēc informācijas atlases principa, kas atrodas automātiska indeksa skanēšanas programmas algoritmā un kataloga darbinieku, kuri atbild par reģistrāciju, uzvedības reglamentā. Parasti tiek salīdzināti divi rādītāji: telpas mērogs, kurā strādā informācijas meklēšanas sistēma (IMS), un tās specializācija.

Sākumā par mērogu. Informācijas masīva formēšanas laikā, meklēšanas sistēma var vērot atjaunošanu vispirms uzdotai dokumentu, katalogu vai mezglu galīga skaita kopai, kas norādīti pēc kāda principa. Tādas sistēmas, kas ir realizētas Internetā, var nosacīti nosaukt par lokālām. Globālās meklēšanas sistēmas, atšķirībā no lokālām, risina darbietilpīgāku uzdevumu – pēc iespējas lielāka visa Tīklā informācijas lauka (WWW, FTP vai cita), ko tās apkalpo, resursu ietveršana. Kā sekas var minēt lomas palielināšanu mehānismam, ko izmanto globālā sistēma, lai palielinātu sev pakļauto mezglu skaitu.

Reģionālu un specializētu meklēšanas servisu veidošana paredz informācijas aktīvu filtrāciju.

Meklēšanas sistēmas specializācija kādas tēmas vai profila bāzē, vai tā būtu cilvēku un organizāciju, datora "dzelžu" vai multimediju failu MP3 formātā , meklēšana, teorētiski var notikt kā lokālā tā arī uz globālā bāzē. Protams, sistēmu ir vienkāršāk veidot un pārskatīt ierobežota atjaunojamo mezglu telpā, ko parasti arī realizē praksē.

Reģionālie meklēšanas dienesti filtrē informāciju galvenokārt uz augstāka līmeņa servera domēna atpazīšanas pamata, piemēram, "lv" Latvijai. Šādu sistēmu nopietns trūkums ir, ka viņi neievēro lielu resursu daļu, kurus izstrādātāji ievieto tradicionāli populārajā domēnā "com".

Reģionālie motīvi bieži tiek ievesti globālo IMS servisā. Lycos sistēma, piemēram, aranžē rezultātus no atsauksmes saraksta atkarībā no tā no kāda reģiona atnāca pieprasījums.

Vēl viens svarīgs virziens meklēšanas servisu reģionalizācijā ir saistīts ar mezglu spoguļu (mirrors) izstrādāšanu populārākām meklēšanas sistēmām. Spoguļiem ir jāsatur pirmatnējā IMS indeksa precīza kopija un jāgarantē pieprasījuma ātra apkalpošana, kas pienāk no noteiktas ģeogrāfiskās zonas. Praksē spoguļu sistēmas indeksa atjaunošana vienmēr notiek ar aizkavi. Tā, AltaVista meklēšanas sistēmas Austrālijas spoguļu sistēmai, līderim pēc spoguļu skaita, tas parasti sastāda 1 – 2 dienas pie bezavārijas darba, un tas ir labākais laiks. Alternatīva starp darba ātrumu un informācijas pilnību kļūst nozīmīga lietotājam, ja viņam ir iespēja vērsties pie spoguļa, un pie pirmavota.

Augstāk jau tika minēts, ka tieši automātisko indeksu, kas aptver noteikta tipa resursus, veidošanai ir zīmju raksturs. Šis notikums vienmēr bija saistīts ar atbilstoša informācijas lauka ātrās attīstības fāzi, bet uz doto brīdi ar WWW telpu. Reāli tikai dokumentu automātiskās indeksēšanas lielais ātrums ar programmu robotu palīdzību var novērst informācijas haosu Tīklā. Taču, pielietojot meklēšanā resursu katalogus "tīrā veidā", bez meklēšanas iespējas pēc

atslēgvārdiem, tā vairāk atgādina sērfingu, nevis nopietnu darbu ar informāciju. Tomēr katalogu loma, kas ir ievērojami kritusi globālā datu uzkrāšanas līmenī, paliek svarīga reģionālā meklēšanā.

WWW katalogi, kas satur loti lielu ierakstu skaitu, piemēram, Yahoo! (vairāk kā 750 tūkst.), nereti izvieto savās lappusēs lokālās meklēšanas mašīnas, realizētas tradicionālo veidņu izskatā. Par cik vizuāli un darbā tās maz atšķiras no veidnēm automātiskajos indeksos, šāda tipa katalogus parasti nepareizi sauc par meklēšanas mašīnām. Būtība nav terminoloģijas pareizībā, kas nav interesanta parastam lietotājam. Problēma ir tajā, ka nesaprašana kā iekšēji funkcionē meklēšanas sistēma, noved pie nekontrolējamas informācijas zaudēšanas. Tā, sekojot kļūdainai definīcijai, var viegli nolikt vienā līmenī globālo automātisko indeksu Northern Light un "meklēšanas mašīnu"- katalogu Yahoo!. Tas nozīmē salīdzināt vienā atslēgā servisus, kas ir paredzēti pavisam dažādu, no profesionālās meklēšanas viedokļa, uzdevumu risināšanai. Kataloga lokālā meklēšanas mašīna paredz meklēšanu pēc atslēgvārdiem, kas ietilpst nodaļu, mezglu nosaukumos un citos nedaudzos datos, kas tiek ievadīti pie reģistrācijas. Tai laikā automātiskā indeksā informācija par noteiktu mezglu ir daudz plašāka, ideālā līdz pat katram atsevišķam vārdam dokumentā, turklāt tiek ievēroti Web lappuses īpašie lauki un datu atjaunošanas režīms.

Lokālās meklēšanas mašīnas Web mezglā organizēšanas vienkāršība, padara to par biežu atribūtu ne tikai katalogiem, bet arī parastiem saitiem. Ja salīdzina lokālās sistēmas indeksa saturu ar informāciju par to pašu mezglu no globālās meklēšanas mašīnas indeksa, tad lokālai sistēmai ir visas iespējas pārsniegt globālo kā pēc datu pilnības, tā arī pēc to atjaunošanas biežuma.

Pateicoties tam, bieži vien visefektīvākais ceļš no pieprasījuma globālā IMS līdz galējam informācijas blokam, iziet caur mezgla lokālā meklēšanas servisa (skat. 2.2.att.) starpposmu. Ar "iekšējo" shēmā tiek saprasta meklēšana gala objekta iekša, ja tas ir iespējams, piemēram, meklēšana Web lappuses tekstā, ko atbalsta lielāka brauzeru daļa.

2.2.att. Meklēšanas procedūras līmeņi.

Ļoti svarīga problēma Tīklā ir dažādu meklēšanas servisu integrācija vienā sistēmā. Tīklam 1999. gads kļuva ievērojams ar vienu nevirzītu notikumu – ar 15 lielāko Interneta

meklēšanas sistēmu dalību, februārī startēja projekts SESP (Search Engine Standards Project), kas tika paredzēts meklēšanas dienestu standartizācijai. Materiālus var atrast pēc adreses http://www.searchenginewatch.com/standards/990204.html.

Jau pirmie dokumenti lika saprast, ka standarta uzdevums ir maksimāli tuvināt sintaksi un meklēšanas valodu iespējas dažādām IMS. Pie tam par vienu no obligātām prasībām kļūst vienotu pieprasījumu komandu atbalsts jebkurā meklēšanas sistēmā, kas lokalizē mezglu pēc tā domēna, bet dokumentu - pēc URL.

Skaidrs, ka pat šī vienkāršā vienošanās paceltu informācijas uzskaiti un kontroli Tīkla mērogā uz principiāli jaunu līmeni.

Teorētiski vilina globālās lieljaudas meklēšanas sistēmas izveidošanas perspektīva, kas varētu sekot Tīklam, tā pilnā informācijas apjomā. Tomēr praksē tas pagaidām nav iespējams, un integrācijas problēmas risinājums nobīdīts uz metameklēšanas sistēmu izveidošanu (skat. 3.1.att.).

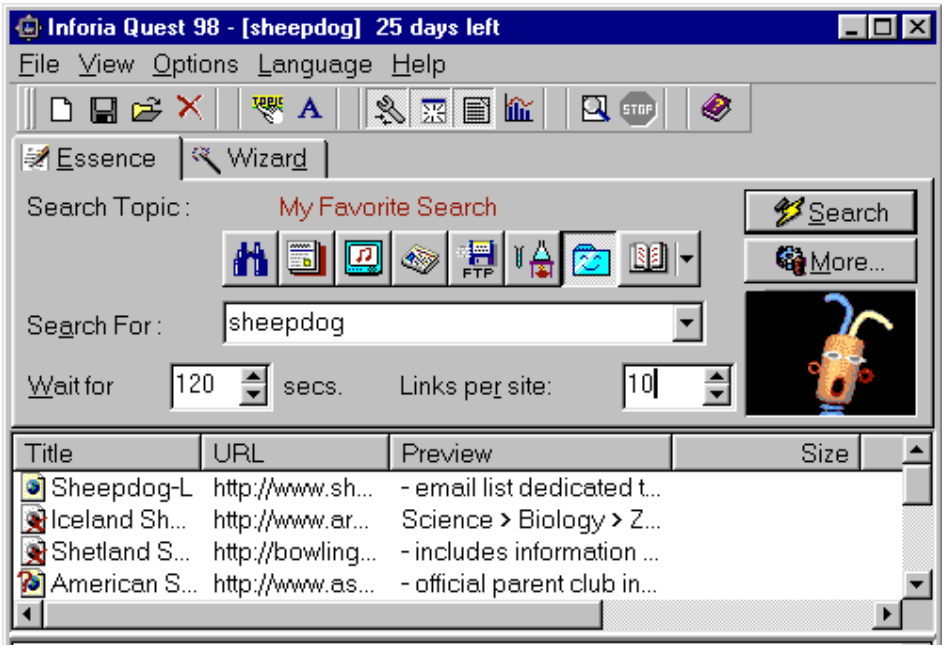
3.2. Metameklēšanas sistēmas

Metameklēšanas sistēma var būt realizēta kā pašā Tīklā, piemēram, uz Telnet jeb Web pieejama mezgla, tā arī lokālas klienta programmas veidā. Kaut arī tai nav pašas indeksa datu bāzes, metameklēšanas sistēma strādā kā vārteja (gateway), kas padod caur savu interfeisu pieprasījumus autonomajām IMS un atgriež meklēšanas rezultātus.

Viens no metameklēšanas servisa uzdevumiem pie meklēšanas – Tīkla testēšana uz informāciju, kas ir relevanta pieprasījumam. Metasistēmas atļauj arī novērtēt atsevišķu IMS pielietošanas efektivitāti pie konkrētā meklēšanas uzdevuma risināšanas. Diemžēl, detalizētai meklēšanai metasistēmas pagaidām maz pielietojamas. Problēma tajā, ka meta vārtejas pieprasījumu valoda ir vispārināta lielākai daļai IMS, un tāpēc sistēmai ir ierobežotas iespējas. Meklēšanas sistēmu standartizācijas projekta SESP parādīšanās šai ziņā paver jaunas perspektīvas metasistēmu attīstībā, par cik IMS standartizācija ievērojami paplašinās vārteju pielietošanas iespējas.

Atzīmēsim, ka metasistēma atļauj pieprasījumu pārsūtīšanu ne tikai uz automātiskiem indeksiem, bet arī tajos katalogos, kas pavadīti ar lokālo meklēšanas mašīnu.

Starp pietiekoši vienkāršiem metameklēšanas lokāliem klientiem, var atzīmēt programmu, kas ir zināma kā Inforia Quest 98 (skat 2.3.att.).



2.3.att. Lokālais metameklēšanas klients Inforia Quest 98

Pēdējās versijas pārbaudes variantu var atrast mezglā http://www.inforia.com.quest.

Pēc pagājuša gada kopsavilkuma, tā tika atzīta par labāko savā klasē un pretendē uz profesionāla meklēšanas instrumenta lomu.

Ātrais šīs programmas pārskats atļauj apzīmēt pēdējās paaudzes metameklēšanas klientu pazīmes.

Pirmkārt, programma sevī ietver ne tikai Web telpas meklēšanas servisus, bet arī citus Tīkla informācijas sektora laukus: failu arhīvus FTP un telekonferences sistēmu.

Pie meklēšanas pieprasījuma apstrādes tiek pieļauts savienojums ar vairāk kā 100 meklēšanas sistēmām, tai skaita arī specializētām.

Atskaites informācija par atrastiem resursiem tiek attēlota programmas darba apgabalā. Norādes, kas dublē jau atrastos, tiek izslēgtas ar sistēmu. Iegūtās adreses uzreiz tiek pārbaudītas uz pieejamību. Ir iespēja izvēlēties nepieciešamo meklēšanas sistēmu kopu no pilna saraksta, uzstādīt meklēšanas laiku un norāžu skaitu, kas tiek saņemts no katra meklēšanas servera, ierobežojumu. Pats IMS saraksts, ar kuru darbojās programma, tiek atjaunots Tīklā automātiski no izstrādātāja servera.

Liels programmas nopelns ir tas, ka tā atbalsta kaut ko līdzīgu meklēšanas valodai: darbojās divi loģiskie operatori un meklēšana pēc frāzēm.

Tomēr katru reizi, kad metasistēmas valoda nevar nodrošināt meklēšanas pieprasījuma precīzu sastādīšanu, nākas izmantot Tīkla autonomos servisus, pirmkārt WWW meklēšanas mašīnas.

4. INFORMĀCIJAS MEKLĒŠANA INTERNETĀ: ZEMŪDENS AKMEŅI

Kas traucē noskaidrot kuru IMS būtu pareizāk izmantot informācijas meklēšanai? Atbilde ir vienkārša: nav pietiekamas informācijas no izstrādātāja puses. Kā sekas tam var minēt saņemamo datu nepatiesumu un to nekontrolējamo zaudēšanu. Reti var atrast Tīklā meklēšanas sistēmu, kurai nepiemistu kādas nedokumentētas īpašības. Liekās, ka lietotājam ir nepieciešamas tikai dažas ziņas: 1)kā notiek IMS datu bāzes uzpildīšana un kāds ir tās apjoms; 2) sistēmas meklēšanas valodas iespēju pilns spektrs; 3)meklēšanas rezultātu attēlošanas pamatīpašības, vispirms atrasto ierakstu aranžēšanas algoritms. Diemžēl šīs informācijas avots parasti ir nevis dokuments, kas pieejams no meklēšanas servera galvenās lappuses, bet gan atsevišķo autoru publikācijas, kas izmētātas pa Tīklu, grāmatām un datoru žurnāliem. Par šādas situācijas iemesliem var minēt ne tikai izstrādātāja paviršību, bet arī faktoru, kas zināms kā marketinga politika. Vienkārši sakot, meklēšanas sistēmas pilnīgākas informācijas attēlošana pati par sevi ne vienmēr pozitīvi attēlojās tās reitingā. Tomēr pārņemt situācijas kontrolei dažos gadījumos ir lietotājam pa spēkam. Noskaidrot izvēlētā meklēšanas servisa darbības īpašības bieži izdodas ar testēšanas palīdzību. Speciālo testu pieprasījumu sastādīšana, kas ātri noskaidro sistēmas darba aspektu nepieciešamā uzdevuma risināšanai, daudzos gadījumos ir netriviāls uzdevums. Tam, kā izvairīties no dažām nepatikšanām strādājot ar IMS, būs veltīts turpmākais apraksts. Kā piemērs tiks izmantotas plaši pazīstamas Interneta meklēšanas sistēmas.

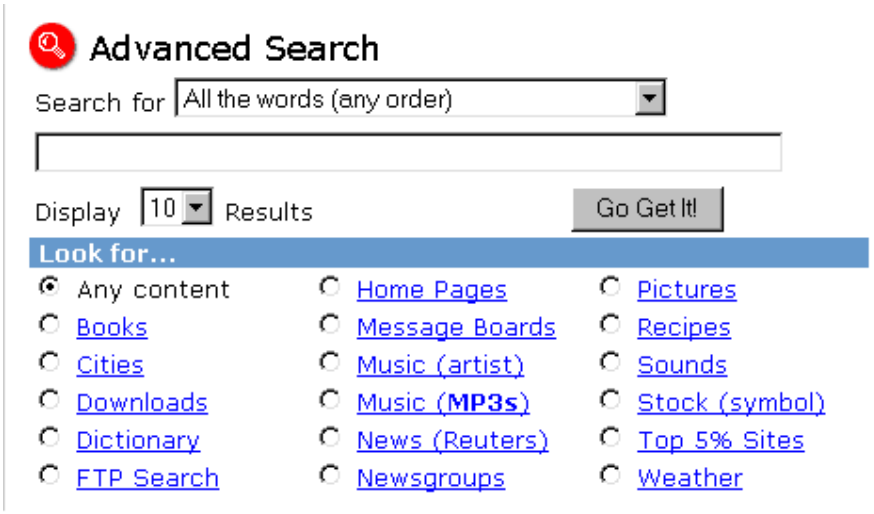
4.1. Problēma N 1: datu bāzes uzpilde

Jebkura meklēšanas mašīna vai katalogs reglamentē savu darbu pēc datu ievākšanas no Tīkla. Acīmredzams, ka informācijas objekta meklēšanas tēla veidošana, vai, citiem vārdiem sakot, tā "attēla" meklēšanas sistēma "spogulī", neatvairāmi ir saistīta ar dažiem kropļojumiem. Būtībā nozīmīgs kļūst jautājums par algoritmu, uz kura pamata veido meklēšanas tēlu. Par objekta oriģinālu turklāt var kļūt kā Web lappuse, tā arī "slēgta" formāta fails, kurš nav pieejams IMS skanējošo programmu piekļūšanai, piemēram, video vai audio ieraksts. Noteiktu veidni pielieto fiziskas personas vai kompānijas meklēšanas tēla veidošanai tās reģistrācijas laikā meklēšanas dienestā. Informācijas filtrēšana no oriģināla piemīt absolūti visām IMS, tai skaitā arī pilna teksta globālas aptveršanas sistēmām.

Filtrācija var reglamentēties kā tehniskajā, tā arī lingvistiskajā līmenī, tomēr uzdevums tai ir viens – ar minimāliem materiāliem tēriņiem panākt meklēšanas reālo efektivitāti.

Sakarā ar to, praksē bieži rodas jautājums, kas kļūst par neveiksmīgas meklēšanas cēloni: pieprasījuma relevantās informācijas trūkums Tīklā uz doto brīdi, vai tas , ka šī informācija potenciāli nav pieejama izskatāmai meklēšanas sistēmai. Par "zemūdens akmeni" šis aspekts kļūst tad, kad ir iegūta kaut kāda atbilde, bet nesaņemto datu daļa kļūst nekontrolējama. Tātad, ja filtrācijas algoritma detaļas nav zināmas, visvairāk jūtami datu zaudējumi parādās pie specializēto meklēšanas dienestu izmantošanas.

Apskatīsim dažus piemērus. Daudzām specializētām sistēmām ir pat savs interfeiss meklēšanas pieprasījumu ievadei. Daudz līdzīgu servisu tiek integrēts globālo IMS veidnēs filtru veidā. Ar tādām īpašībām ir slavens HotBot, nesen līdzīgas iespējas ieviestas AltaVista, ir tie arī uz Excite. Pastāvīgi paplašinās meklēšanas sistēmas Lycos filtru kopums (skat 3.1.att.), kuru izskatīsim rūpīgāk.



3.1.att. Paplašinātas meklēšanas veidne Lycos ar vairāku filtru atbalstu.
(http://lycospro.lycos.com/)

Iedomājaties sevi lietotāja vietā, kurš pirmo reizi ir atnācis uz tik pazīstamu globālo meklēšanas sistēmu kā Lycos, ar velēšanos atrast Tīklā ziņas par kādu grāmatu. Ievadot atbilstošos atslēgvārdus un izvēloties filtru "Books", mēs iegūstam atsauci, ko pie papildus informācijas trūkuma nevar novērtēt savādāk, kā datu ievākšanu par grāmatām, savāktām no visa Interneta. Būtu interesanti uzdot jautājumu, vai Tīkla mērogā ir iespējama automātiska tādu ziņu ievākšana? Ja runā tikai par WWW telpu, tad vairākumā gadījumu programmas zirnekļi, skenējot Tīklu, izmanto datu tipa atpazīšanai HTML valodas speciālos elementus, ar kuru palīdzību Web lappusē tiek ievesti noteikti informācijas bloki. Elementa nosaukumam var būt nozīmes slodze, tas var tikt salīdzināts ar informācijas tipu. Tā, ja hipotētiski pastāvētu HTML elements "book", kas sevī ietvertu ziņas par grāmatu un autoru, tas varētu tikt izvietots lappusē un tam būtu tādā veidā: <book> Grāmatas nosaukums un autors</book> (pašam elementam

<book> vārtejas logā nav jāattēlojas). Turklāt visa informācija par grāmatām, kas tiek publicētas šādā veidā WWW, varētu veiksmīgi arī bez cilvēka līdzdalības uzkrāties IMS datu bāzē. Bet "book" elements pašlaik HTML standartā nepastāv. Tādēļ nākas izmantot vai nu manuālo atlasi, vai automātisko apskati dažu mezglu iepriekš uzdoto katalogu, kam iespējams ir sakars ar grāmatu tirdzniecību vai bibliotēkām.

Lycos gadījumā viss ir daudz vienkāršāk. Meklēšana notiek tikai pēc viena vienīga kompānijas mezgla (www.barnesandnoble.com), kura ir ieinteresēta savas preces realizācijā. Izstrādātajam par godu jāsaka, ka pēc vairāku gadu klusēšanas par filtru "books", šodien piedāvātas dokumentācijas dziļumos var atrast nelielu ieminēšanos par filtra nomnieku. Agrāk tā īpašnieku vienkārši nebija iespējams identificēt, un tikai pēc kāda laika kļuva skaidrs, ka sistēma strādā ar pietiekami neievērojamu pēc tilpuma un specifiski piepildāmo datu bāzi.

Ne mazāk nopietnas izskatās bažas, kad meklēšana ir saistīta ar informāciju, kas ir saistīta ar tās glabāšanas noteikto formātu, piemēram, skaņas failiem. Dažu mēnešu laikā "skaņu Internetā" meklēšana uz Lycos kļuva par kaut ko noslēpumainu, kas atgādināja darbu ar nelielu, bet ar gaumi atlasītu kolekciju. Sistēmas testēšana ar parasto pieprasījumu palīdzību rādīja, ka pamatā tajā ir atveidoti formāti wav un au. Nesen kļuva zināms, ka tagad tiek atbalstīti arī formāti mp3, mid, ra, ram un aif. Pie tam uzkrāto ierakstu apjoms, kas ir pieejams caur filtru vairākumu, tiek turēts slepenībā.

Ir skaidrs, ka, ja jūs interesējošais formāts pašlaik nav sistēmas atbalstāmo sarakstā, tad jūs saņemsiet nulles atsauksmi, kā iemeslu varēja paredzēt jau no paša sākuma.

Pavadrakstu rašanās skaņas failiem uz Lycos, kas parādās, attēlojot meklēšanas rezultātu, joprojām nav reglamentēti izstrādātājam.

Analoģiskās problēmas pastāv arī citās IMS. Gribas atzīmēt tipisko paņēmienu: globālo IMS veidņu izmantošana informācijas meklēšanai kā visā Interneta telpā, tā arī meklēšanai pēc dažām izvēlētām datu bāzēm un kolekcijām. Diemžēl reālais meklēšanas lauks ne vienmēr tiek iepriekš aprunāts, un bieži vien to nākas noskaidrot pašam, lai neizdarītu nepareizus secinājumus turpmāk.

4.2. Problēma N 2: meklēšanas pieprasījumu valoda

Situāciju var sarežģīt tas, ka meklēšanas serverī jūs neatradīsiet izsmeļošu informāciju par to kā strādā pieprasījuma valodas operatori. Pat jau nobriedušās, ne pirmo gadu strādājošās IMS, ar to var sastapties. AltaVista piemērā izskatīsim, kā tas var kļūt par noteiktu problēmu avotu.

Neskatoties uz jauna grafiskā filtra parādīšanos (skat .2.att.), daudzi sistēmas lietotāji turpina izmantot pēc savas būtības vienkāršu operatoru "image", kas ļauj atrast indeksā grafiskos failus. AltaVista izziņa rekomendē ievadīt veidnē pieprasījumu, kur aiz norādītā operatora ir jābūt meklētā faila nosaukumam vai nosaukuma daļai. Tādā veidā, lai atrastu failu ar akropoļu attēlu, ir jānorāda pieprasījums veidā "image: acropolis".

3.2.att. AltaVista vienkāršās meklēšanas veidne (www.altavista.com) ar filtriem un meklēšanas valodas izvēlnēm.

Vai zināšanas par to kā reāli strādā operators "image" palielinās mūsu izredzes uz veiksmi? Ja paskatīsimies uz atsauktiem dokumentiem, un pēc tam uz to HTML avotu, tad viegli pārliecināsimies, ka katrā no viņiem grafiskā attēla vietā atrodas elements <img>. Tā iekšā kā obligāts atribūts ir URL, no kura arī tiek izsaukts fails: <img src = "http://www.citforum.ru/buildings/acropolis.gif">

Faktiski Web lappuse dod atsauci, ja atslēgvārds ietilpst ne tikai faila nosaukumā, bet arī ja to satur jebkura kataloga nosaukums un servera domēna vārds, kas ir norādīti <img> elementa URL. Tas ir dokuments, kas ietver sevī agrāk noradītu rindiņu, atsaukto uz pieprasījumu "image: buildings". Tātad, meklēšana pēc kataloga vārda, kurš tāpat kā faila vārds nes nozīmes slodzi, atļauj iegūt grafiskos datus, kurus nevar dabūt pirmajā gadījumā. Pieņemsim, ka Web meistars neuzmanības dēļ nosauca meklēto failu acr1.gif, bet pareizi ievietoja to katalogā buildings. Tad pēc pieprasījuma "image: buildings" var atsaukties relevanti dokumenti ar akropoļu attēlu, kas ir ievietoti Web lappusē ar rindiņu <img src = "http://www.citforum.ru/buildings/acr1.gif">

AltaVista paplašinātā meklēšanā tiek izmantoti loģiskie operatori un iekavas. Tomēr uz servera nav nekas teikts, vai ir atļauta to pielietošana speciālo meklēšanas laukumu iekšienē, tādos kā laukums "image". Jau iepriekš piereģistrētā indeksa grafisko failu, kas tika atrasts agrāk, var izmantot atsevišķo meklēšanas pieprasījumu pārbaudei. Tā, ja pieņem, ka fails ar URL no pēdēja piemēra, pastāv, tad testa pieprasījumam "image: (buildings AND acr1) veidā, ir jābūt

kaut kādai atsaucei, un tādā veidā jāapstiprina, ka operatoru kombinēšana ir atļauta. Praksē tas tiešām ir iespējams.

Gribētos vēlreiz pasvītrot, ka runa iet nevis par atsevišķo meklēšanas sistēmu izturību, bet par konstruktīvu jautājumu risināšanas pieeju. Pie tam nav retas situācijas, kuras paredzēt ir ļoti grūti.

Populārā biznesa – orientētas sistēmas Open Text Livelink Pinstripe (OTPL) paplašinātas meklēšanas veidnē (skat 3.3.att.) arī ir paslēptas dažas problēmas, kuras nav aprakstītas IMS izziņu materiālā.

3.3.att. OTPL paplašinātas meklēšanas veidne.
(http://pinstripe.opentext.com/search/power.html)

Kā var redzēt no attēla, veidne ļauj uzdot savu meklēšanas lauku katram terminam, un pēc tam saistīt terminus ar loģisko operatoru palīdzību. Tomēr, kad terminu skaits kļūst lielāks par divi, rodas jautājums, kādā secībā tiks apstrādāti operatori un kāds būs rezultāts. Pat tik vienkāršam pieprasījumam kā "term1 AND term2 OR term3" ir prātīgi piedāvāt dažādu interpretāciju, ko var ilustrēt ievietojot iekavās loģiskās vienības (pašā veidnē iekavas netiek pielietotas). Gan variants "(term1 AND term2) OR term3", gan variants "term1 AND (term2 OR term3)" liekās pieņemami, dodot rezultātā absolūti dažādas atsauces. Testa pieprasījums un turpmākā iegūto dokumentu analīze parāda, ka izpildās pirmais variants, t.i. ka operatori izpildās parādīšanās kartībā veidnē, un dokumentā būs vai nu "term1" un "term2" vienlaicīgi, jeb tikai "term3".

Interneta IMS lielākā daļa šodien aktīvi strādā ar tā sauktiem stop vārdiem (stop – words). Pie pēdējiem pieder daži Tīklā visbiežāk lietotie vārdi kā information, Internet, Web, business un citi. Ir zināms, ka AltaVista, Excite, HotBot un Lycos pielieto darbā stop vārdu tehniku, bet InfoSeek un Northern Light to nepraktizē.

Parādoties stop vārdiem meklēšanas pieprasījumā bez speciālām viltībām, IMS var tos neievērot pie meklēšanas un rezultātu aranžēšanas, dažreiz informējot par to lietotāju, dažreiz – nē. Kopumā stop vārdu neievērošana pie pieprasījuma apstrādes, samazina meklēšanas laiku un palielina atsauces relevanci. Tomēr, ja jūs gribēsiet atrast kaut ko līdzīgu klasiskai Šekspīra frāzei "to be or not to be", kas sastāv tikai no stop vārdiem, jūs vairs nepārvaldāt situāciju.

Kaut arī stop vārdi var tikt ignorēti parastajos pieprasījumos, pilna teksta IMS indeksā viņi pastāv tā pat kā pārējie. Tāda sistēma, piemēram, ir AltaVista (tiek indeksēti visi dokumenta vārdi). HotBot savukārt pilda pretējo, indeksē visu izņemot stop vārdus.

Bet tomēr arī HotBot izpilda atsevišķo nozīmīgo dokumenta laukumu pilna teksta indeksēšanu, tādēļ pieprasījumi ar stop vārdiem, kas ir noformēti frāzes veidā, dod šajā IMS rezultatīvo atsauci.

Stop vārdu saraksts nav standartizēts, tāpēc tas var būt oriģināls katram servisam. Izstrādātāji reti dod ziņas par šo IMS darbības aspektu, tomēr, ja ir nepieciešams, meklēšana pēc vārdiem "stop", "words" plus jūs interesējošās meklēšanas sistēmas nosaukums ļauj atrast Tīklā atbilstošo sarakstu versijas.

Vispārināti principi, lai izietu no problēmas situācijas, ir sekojošie: pēc iespējas izvairīties no stop vārdu izmantošanas pieprasījumos, izslēgt AND, OR, NOT un citu loģisko operatoru izmantošanu tajās veidnes, kurās tos neatbalsta un tiks uztverti kā stop vārdi.

Ja bez stop vārdiem nevar iztikt, tad vajadzētu tos ieslēgt frāzē, kas vairākās sistēmās nozīmē ievešanu pēdiņās. Atsevišķos gadījumos ir lietderīgi notestēt parastās un paplašinātas meklēšanas IMS veidņu darbību, kuros stop vārdu atbalsta tehnika varbūt dažāda.

4.3. Problēma N 3 : meklēšanas sistēmas atsauce

Pati interesantāka intriga, ko rada IMS, ir saistīta ar algoritma, kas aranžē rezultātus atsauces sarakstā, darbības īpašībām. Šīs ziņas parasti netiek plaši izpaustas, bet tās ir nepieciešamas Web meistariem, kuri popularizē bargā konkurences cīņā savus mezglus caur Interneta meklēšanas sistēmām. Iekļūt pirmajos dažu desmitu ierakstos no IMS atsauces saraksta pēc bieži atkārtotiem Tīklā pieprasījumiem nozīmē nodrošināt savu pieejamību potenciāliem klientiem.

Tomēr risinot meklēšanas uzdevumu, darbā ar atsauces sarakstiem arī var rasties problēmas, informācijas trūkuma dēļ.

Ne katra IMS dod pilnu dokumentu sarakstu, kas ir atsaucē uz pieprasījumu (piemēram Lycos nedod). Zināmā mērā tas ļauj sistēmai saglabāt savu seju, izvairoties no salīdzinājuma ar

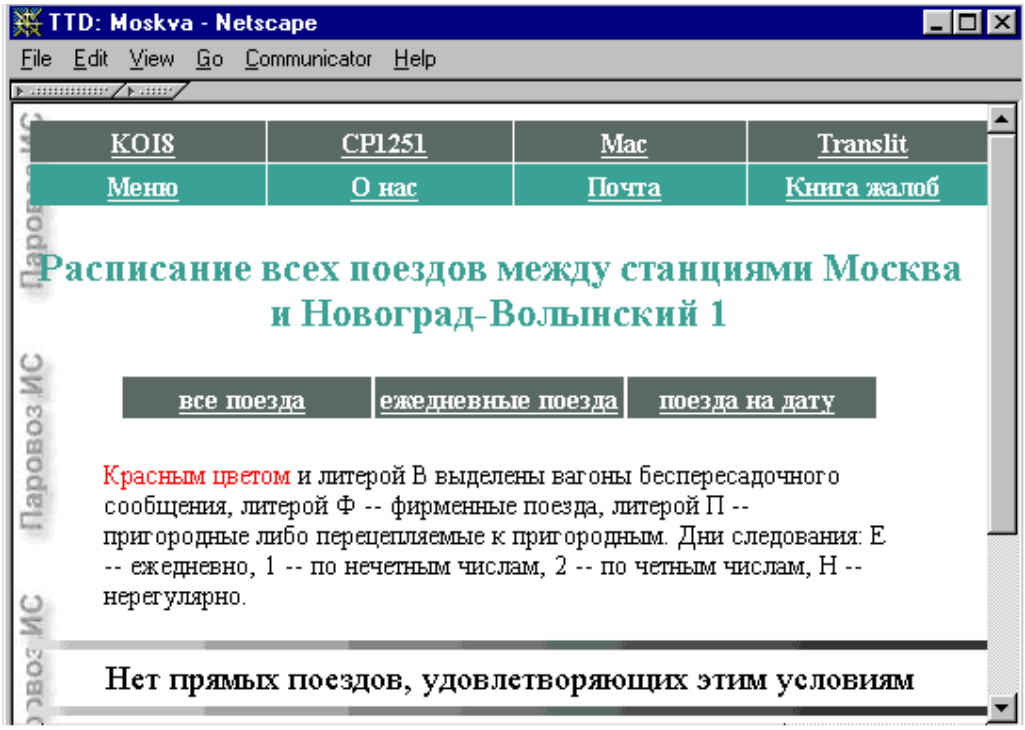
gigantiem – Northern Light, AltaVista vai HotBot. Profesionālo uzdevumu risināšanā pie tādiem servisiem ir jāgriežas pēdējā gadījumā.

Parasti atsauces sarakstā parādās informācija, kura ietver sevī lappuses virsrakstu, adresi un anotāciju. Anotācija ņem vai nu no speciālā META elementa, kuru uzdod dokumenta autors, vai nu par to pieņem no lappuses pirmās dažas nerediģējamas teksta rindas. Dažos gadījumos norāda dokumenta valodu. Vēl var gadīties, ka atrastos dokumentos nav atslēgvārdu, pēc kuriem tika organizēta meklēšana. Par tādas parādības iemeslu, neskaitot lappuses nereģistrēto atjaunošanu bez adreses izmaiņas, var būt fakts, ka atslēgvārdus uzstādīja autors speciālā laukā META elementā. Pēdējais pieejams IMS robotam skanēšanai, bet netiek attēlots lappusē. Šai gadījumā ar META elementu apskati HTML avotā ir iespējams pārliecināties par autora neapzinīgumu: atslēgvārdu neatbilstība dokumenta saturam ir tīša dezinformācija.

Vēl viena problēma lietotājam vispār nav acīmredzama. Runa iet par to, kā meklēšanas serveris apstrādā pieprasījumus gadījumā, kad to ir pārāk daudz, t.i. pārpildes režīmā. Tā, piemēram, uz AltaVista, kad ir pienākuši praktiski vienlaicīgi vienādi testa pieprasījumi no 10 - 15 datoriem, rezultātu skaits, kas paradās atsauces sarakstā katram sistēmas lietotājam dažreiz var atšķirties par desmitiem tūkstošiem. Īstenībā, nokļūstot pārslodzes režīmā, meklēšanas serverim nav lielas izvēles: vai nu atraidīt pieprasījumu, vai nu apkalpot to pēc "saīsināta" varianta. Pēdējais var piedāvāt atrasto datu, kas apmierina pieprasījumu, daļas attēlošanu. Izeja ir acīmredzama: IMS atsauces patiesīgums ir jāpārbauda daudzkārtēji un dažādā diennakts laikā.

4.4. Problēma N 4 : nevīžība un mistifikācijas

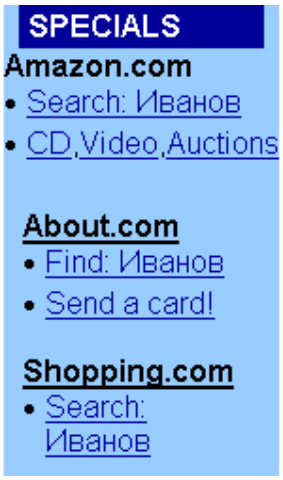
Šeit gribētos apstāties uz dažām vairāk kā reālām briesmām, kas sagaida lietotāju, kurš paļāvās uz maz pazīstamu meklēšanas serveri. Ir zināms viens gadījums. Cilvēkam bija nepieciešama steidzama informācija par tiešo elektrovilcienu esamību starp divām NSV pilsētām. Izmantojot Rambleru, viņam ātri izdevās lokalizēt serveri, kas piedāvāja nepieciešamās ziņas (skat.3.4.att.).



3.4.att. Meklēšanas pieprasījuma apstrādes rezultāts uz "pilnā dzelzceļu saraksta" servera pa Krieviju un NSV

Pēc atiešanas un gala staciju ievadīšanas, sistēma atbildēja negatīvi (attēla apakšējā rindiņa: "Nav tiešo vilcienu, kas atbilst šiem noteikumiem"). Tāda kategoriska servera atbilde piespieda cilvēku pārtraukt turpmāko meklēšanu un pieņemt lēmumu, kuru drīz vien nācās nožēlot. Iesniegt pretenzijas sistēmas izstrādātājam arī nebija iespējams. Lieta ir apstāklī, ka uzreiz zem meklēšanas rezultāta lietotājs nepamanīja vienu svarīgu detaļu – uzrakstu "Reklāmas saraksts, iespējamas izmaiņas, par kurām nenes atbildību ne izplatītājs, ne IMS". Turklāt, ja atteices frāze būtu noformulēta "mīkstāk", lietotājs, iespējams, varētu turpināt meklēšanu Tiklā un iegūt pozitīvu rezultātu.

Dažos gadījumos izstrādātāja marketinga agresivitātei ir izsaucošs raksturs. Jau vairākus mēnešus HotBot un AltaVista serveros izvietoti lielākās grāmatu tirdzniecības kompānijas Amazon (www.amazon.com) un citi reklāmas paziņojumi. Pie tam uz jebkuru pieprasījumu, IMS blakus meklēšanas rezultātiem parādās banners, kas dod mājienu uz to, ka informāciju, tieši pēc izpildītas meklēšanas tematikas, var atrast uz Amazon, pat ja pieprasījumā figurēja mistiskais "Ivanova kungs" (skat. 3.5.att.).



3.5.att. "Gudrais" banners uz servera www.altavista.com

Terminu ievietošana no meklēšanas veidnes bannerī tiek izpildīta ar to mehānisko pārvietošanu, bez jebkādas kontroles par šīs tematikas grāmatu reālo esamību uz kompānijas servera. Turklāt atrast "Ivanovu" uz Amazon nav iespējams principā, par cik līdz šim momentam literatūra krievu valodā tur netika pārdota. Šai gadījumā maksa par uzticēšanos – dažas minūtes veltīgi iztērēta laika.

Tāpēc no parastās cieņas pret grāmatām Tīklā labāk ir atteikties, īpaši ja serveris ģenerē replikas automātiski.

KOPSAVILKUMS

Viena no 20.gadsimta beigu pazīmēm ir pasaules vienotas informatīvās telpas veidošanās process. Tā viena no spilgtākajām ārējām izpausmēm ir pasaules globālais datoru tīkls internets. Internets ir gan lielisks tehnisks sasniegums, gan stingru noteikumu, likumu un izvēles brīvības sadzīvošanas vieta. Intelektuālais ieguldījums no visas pasaules veido interneta bagātību, un interneta demokrātiskais raksturs garantē mums iespēju papildināt šīs bagātības. Daudziem no mums internets ir kļuvis par vienu no svarīgākajiem informācijas avotiem, daudziem tas ir pats galvenais un vienīgais. Internets ir milzīga laboratorija, kurā tiek pārbaudītas visneiedomājamākās idejas un risinājumi. Tomēr, darbojoties internetā, būtu jāizvērtē atrastās informācijas kvalitāte un patiesības pakāpe.

Interneta mīnusi -

- Informācija internetā ir vāji organizēta (internets ir kā milzīga nesakārtota bibliotēka, kur vajadzīgā informācija jāatrod bez bibliotekāra palīdzības);
- Internetā pieejamā informācijas ir ar nezināmu kvalitāti;
- Internets ir nedrošs (internetu izmanto visi sabiedrības slāņi, tajā parādās visai sabiedrībai kopumā raksturīgās pozitīvās un negatīvās iezīmes);
- Interneta piekļuves tehnoloģijas nav ergonomiskas (ne vienmēr ir pieejams tur un tad, kad tas ir vajadzīgs)
- Interneta lietošana nav atpūta (meklējot informāciju internetā, ir jākoncentrējas, nepārtraukti rodas izvēles situācijas);
- Interneta lietošana prasa daudz laika (pirms nonāk pie meklētās informācijas, bieži jāveic ilgstoši meklējumi).

Informācijas sistēma ir ļoti ietilpīgs termins, ar kuru tiek identificēts viss, kas nodrošina konkrētu informācijas organizāciju. Informācijas sistēmai nav noteikti jābūt balstītai informācijas tehnoloģijās, tomēr diez vai papīra kartīšu kartotēku var uzskatīt par modernu, efektīvu un informatīvo vajadzību apmierinošu. Informācijas sistēmām jābūt elastīgām, viegli izmaināmām un ātri jānes augļi. Tām ir jādzimst, lai mainītos. Informācijas meklēšanas sistēma ir līdzekļu un metožu kopums, kas paredzēts dokumentu, ziņu par dokumentu, faktu glabāšanai un meklēšanai atbilstoši informācijas lietotāja nepieciešamībai. Informācijas sistēmas paredzētas datu apstrādei, kantoru darbu un ekspertu sistēmu uzdevumu automatizācijai. Tā ietver aparatūru, programmatūru, informāciju un lietotāju. Visas informācijas meklēšanas sistēmas ir izveidotas ar mērķi:

- vākt
- glabāt
- apkopot
- sistematizēt
- piedāvāt

informāciju.

Informācija ir jebkuras informācijas sistēmas darba rezultāts.

Starptautiski pazīstamākās informācijas meklēšanas sistēmas (search engines) ir AltaVista, Excite, InfoSeek, Netscape, Lycos, Looksmart, Dogpile, Go TO, Yahoo. Šīs minētās Web vietas ir labākās, ja jāmeklē informācija par kādu vispārīgu tēmu.

AltaVista (www.altavista.com) ir spēcīgs vispārīgs meklēšanas rīks, kuru kā vienu no pirmajiem meklētājiem internetā ieviesa pazīstamā datorfirma Compaq. Kaut arī AltaVista nepiedāvā dažus meklēšanas nosacījumu veidošanas rīkus, kādi ir citiem meklētājiem, šeit ir atrodamas dažas unikālas iespējas, piemēram, iespēja pieprasījuma lauciņā ievadīt vienkāršus jautājumus. Šobrīd AltaVista sistēmā ir iespējams strādāt 25 valodās. Ir iespējama vārdu, frāžu vai pat visas tīkla lappuses tulkošana no vienas valodas citā. Šajā tulkā pagaidām nav iekļauta latviešu valoda. Meklēšanas servisu ir viegli lietot – jāievada tikai jautājums, atslēgas vārds vai frāze, jānoklikšķina "search" vai jānospiež "enter". Rezultāti tiek sakārtoti tā, lai labākie varianti atrastos saraksta sākumā. Visvairāk meklētajām lietām meklēšanas serviss ir jau sagatavojis standartus, kas atvieglo meklēšanu. Tie parādās AltaVista meklēšanas sistēmas pirmajā lapā ar nosaukumu "kategorijas". Ja nepieciešams izmantot Bula operatorus, tad jāizmanto paplašinātās meklēšanas serviss (advanced search service). AltaVista piedāvā ātru un ērtu meklēšanas servisu ar vairāk nekā 125 000 WWW lappusēm. Ir iespēja meklējumus sašaurināt, tāpat arī iespējams meklēšanas rezultātus attīrīt, lai iegūtā informācijas būtu pēc iespējas precīzāka un atbilstošāka. Lai uzlabotu meklēšanas rezultātus, AltaVista piedāvā vairākus rīkus meklējumu uzlabošanai, piemēram, pēdiņas, loģiskos operatorus + un – , zvaigznīti, kas var aizstāt vairākus simbolus. Vēlams lietot tikai mazos burtus, jo tad tiks meklēti gan ar mazajiem, gan lielajiem burtiem rakstīti vārdi. Vārdus pēdiņās AltaVista saprot kā frāzes. Lai sameklētu informāciju kādā noteiktā valstī vai kādā konkrētā Web vietā, aiz "host" jāievada URL daļa, kas norāda valsti un meklējamo vārdu. Šeit gan jāzina valstu saīsinājumi. Lai sameklētu attēlus, aiz vārda "image" ievada attēla nosaukumu.

Excite (www.excite.com) – veicot meklējumus šeit, iespējams iegūt rezultātus ne tikai no globālā tīkla, bet arī no ziņām, enciklopēdijām, skaņu failiem. Excite ir brīvs, personisks

onlaina serviss tīklā. Kas piedāvā 16 programmētus tīkla kanālus, privāto e-pastu un onlaina iepirkšanos.

Excite patentētā ICE meklēšanas tehnoloģija dod iespēju lietot vairāk kā 50 miljonus Web lappušu, 140 000 iepriekš atlasītus Web lappušu sarakstus un 1000 Usenet interešu grupas. Šeit sistēma sameklē dokumentu ne tikai pēc ievadītā atslēgas vārda, bet arī radniecīgus pēc satura. Excite piedāvā arī divus pakalpojumus, kas dod iespēju padarīt meklējumus precīzākus:

- search wizard piedāvā terminus;
- power search veic meklēšanu pat tad, ja lietotājam nav zināšanu par Bula operatoriem.

Excite piedāvā izmantot iespēju piedalīties diskusiju grupās (Excite Boards un Excite PAL) un arī onlaina diskusijās. Tāpat ir iespēja izmantot Fast Facts, kas satur uzziņu saites ("dzeltenās lapas", personu meklētājs, vārdnīcas).

Loģiskie operatori, Bula operatori u.c. līdzīgi kā AltaVista sistēmā.

InfoSeek (www.infoseek.com) kā viens no komponentiem ietilpst Go Network sastāvā, kas savukārt saistīts ar Disney un ABC Web lapu tīkliem. Tāpat kā lielākajai daļai meklētāju, arī InfoSeek, pamatmeklēšanas forma sastāv no vienas rindas. Līdzīgi kā HotBot un Lycos, šeit ir arī izvērstā meklēšanas forma, kurā var veidot loģiskas izteiksmes, kā arī meklēt pēc kategorijām.

Lycos (www.lycos.com) ir ērts un pievilcīgs meklētājs, tas meklējumu rezultātu sarakstos iekļauj ziņas, piedāvājumus, ir iespēja meklēt tīklā, personālajās mājas lapās, tīkla lappušu paskatā, ziņās un citos apgabalos. Lycos Pro Search lapas piedāvā meklēt specifisku informāciju, piemēram, attēlus, mp3 failus, grāmatas, ziņu grupas.

Tāpat kā citās sistēmās arī šeit var izmantot loģiskos operatorus – plus un mīnus, lai attiecīgi izslēgtu vai pievienotu kādu vārdu, Bula operatorus. Web lappušu sarakstu, kas sameklēts pēc lietotāja pieprasījuma, sauc par meklēšanas rezultātu lappusi (Search Results Page). Šeit lietotājs var atrast:

- sakārtotas tīkla lappuses (Matching Web Pages), kur lappuses centrā atrodas tīkla lappušu saraksts, kas satur meklētajai informācijai visatbilstošākos rezultātus. Rezultāti sagrupēti, ir adreses vairākām mājas lapām;
- nākošā lappuse (Next Page) vai iepriekšējā lappuse (Previous Page) palīdz lietotājam orientēties daudzajās lapās. Saraksta apakšā ir arī lappušu numuru indekss, kas ļauj nokļūt jebkurā lapā;
- radniecīgo tematu pārskatā (View Related Topics) piedāvā tīkla lappuses, kas satur pieprasītajam atslēgas vārdam vai frāzei radniecīgu informāciju
- Pictures&Sounds piedāvā informāciju par multimedijiem.

Tāpat kā visas informācijas meklēšanas sistēmas, arī Lycos piedāvā palīdzību, kā efektīvāk veikt meklēšanu, lai gan lielākajā daļā informācijas sistēmu meklēšanas iespējas ir līdzīgas, daudzās pat identiskas, atšķiras informācijas kvalitāte un kvantitāte.

Yahoo (www.yahoo.com) – šis ir ne tikai viens no labākajiem, bet arī pats populārākais tīklā. Yahoo struktūra ir pārdomāta, rūpīgi izstrādāta un tiek pastāvīgi uzturēta, papildināta, attīstīta. Yahoo ir ideāli piemērots sākotnējiem meklējumiem, tas piemeklēs meklēšanas kritērijam atbilstošas kategorijas un atbilstošas lapas. Ērti lietojams ir Yahoo kategoriju saraksts, ir iespēja meklēt tikai konkrētas kategorijas ietvaros.

Ja nepieciešams atrast kādu specifisku informāciju, tad labākus rezultātus dod meklēšana vietās ar konkrētu specializāciju – diskusiju grupas, ziņas, informācija par kompānijām, publicēti raksti u.c., nevis meklēšana AltaVista vai Yahoo.

LITERATŪRA

1. Informācijas meklēšanas sistēma AltaVista: Informācija par serveri / Internets. - http://www.altavista.com/

2. Informācijas meklēšanas sistēma Yahoo: Informācija par serveri / Internets. - http://www.yahoo.com/
3. Informācijas meklēšanas sistēma Lycos: Informācija par serveri / Internets. - http://www.lycos.com/
4. Informācijas meklēšanas sistēma Excite: Informācija par serveri / Internets. - http://www.excite.com/
5. Sataki K. Kā tīklā veiksmīgi sameklēt informāciju. – Rīga: Datorpasaule, 1999. Septembris. 50.-52.lpp.
6. Misa R. Internets un ceļojumi. – Rīga: Datorpasaule. 1997. Februāris. 20.-22.lpp.
7. Anspoks A. Informācijas sistēmas 21.gadsimtam. – Rīga: Datorpasaule. 1999. Oktobris. – 33.-36.lpp.
8. Rakstu krājums par meklēšanu internetā / Internets. - http://www.citforum.ru/pp/
9. Norādes uz rakstiem par meklēšanas sistēmām un iespējam / Internets. - http://www.zhurnal.ru/search/articles.shtml
10. Raksti par meklēšanas iespējām / Internets - http://www.zdnet.com/products/internetuser/search.html