Tehnoloģiskie risinājumi interneta meklētājiem

GINTS ERNESTSONS

Informācijas meklējumiem internetā Latvijas datorlietotāji šobrīd var izmantot veselu virkni dažādu meklētājserveru. Populārākie no tiem Latvijā ir Lursoft *(www.lursoft.lv)* un Apollo *(www.apollo.lv)* portālu izmantotais SIETS.LV meklētājs *(www.siets.lv),* kā arī portāla TvNet *(www.tvnet.lv)* izmantotais LATNET meklētājs. Salīdzinoši nesen sākuši darboties arī Delfi SMART *(smart.delfi.lv)* un Tildes LETONIKA *(www.letonika.lv)* interneta resursu meklētāji. Pasaulē pazīstamākie interneta meklētāji ir Google *(www.google.com)* un Yahoo *(www.yahoo.com),* tagad tiem pievienojies arī Microsoft MSN meklētājs *(www.msn.com).*

Neskatoties uz visai askētisko lietotāju interfeisu, šo serveru lietotāju skaits ir pārsteidzoši liels. Tomēr šis pieticīgais interfeiss ir visai mānīgs. Tas ir kā aisberga redzamā daļa un patiesībā tehnoloģiskais risinājums jebkurai šādai sistēmai nebūt nav triviāls.

IZAICINĀJUMI INTERNETA MEKLĒTĀJU VEIDOTĀJIEM

Veidojot interneta meklētājprogrammatūru tehnoloģiskos risinājumus, pirmā problēma, ar kuru nākas saskarties, ir lielais nestrukturēto datu apjoms. Internetā publicētās informācijas indekss pat nelielas valsts ietvaros var aizņemt simtus gigabaitu, nereti – vairākus terabaitus. Tradicionālās uz SQL relāciju datu bāzēm bāzētās sistēmas nav piemērotas nestrukturētu datu ātrai apstrādei, kas ir obligāts priekšnoteikums kvalitatīvam meklēšanas pakalpojumam.

Otra problēma ir risinājuma tehnisko līdzekļu izmaksas šo datu uzglabāšanai un meklēšanas indeksa izveidošanai. Meklēšanas pakalpojumu sniedzējam ar milzīgo datu masīvu ir jāstrādā ikdienā, šāds interneta indekss ir regulāri jāpapildina un jāaktualizē, nepārtraucot meklēšanas servisu lietotājiem. Tādēļ gan nacionāla, gan globāla mēroga meklētājus nav iespējams kvalitatīvi izveidot kā sistēmas, kas darbojas tikai uz viena paša servera. Lai cik jaudīgs nebūtu viens serveris, tradicionālajā PC serveru arhitektūrā ir limitēts gan vienam serverim pieslēdzamo disku maksimālais skaits, gan operatīvās atmiņas daudzums. Pilns interneta indekss viena servera atmiņā nevar brīvi ietilpst, bet tas ir obligāts priekšnoteikums servisa ātrdarbībai. Daudzi grēko ar kompromisa risinājumu uz ātrdarbības rēķina, taču tad, palielinoties lietotāju skaitam, pakalpojumu kvalitāte būtiski pasliktinās un izraisa neapmierinātību lietotājos.

Trešā problēma ir metodika, kādā veidā savākt sākotnējo datu kopumu indeksa izveidošanai un pēc tam sakārtot meklēšanas rezultātus tā, lai lietotāji būtu apmierināti un uzskatītu tos par pietiekami labiem.

KĀ UZBŪVĒTS TIPISKS INTERNETA MEKLĒTĀJS

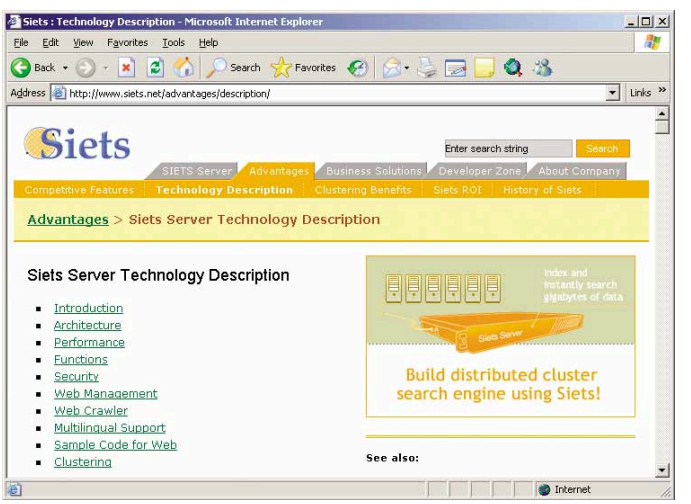
Darbības principu ilustrācijai izmantosim tehnoloģiju Siets, jo pēc līdzīgiem principiem ir veidota lielākā daļa visu pārējo interneta meklētāju tehnoloģiju, arī, pasaulē vadošās – Google un Yahoo.

Interneta meklētājs Siets.lv izmanto tāda paša nosaukuma meklēšanas tehnoloģiju, kuras apraksts atrodams *www.siets.net* (sk. 1. attēlu). Šobrīd tā ir vienīgā meklēšanas tehnoloģija Latvijā, kuras programmatūru var brīvi lejupielādēt izmēģināšanai un izmantot dažādos – gan Interneta, gan korporatīvos risinājumos.

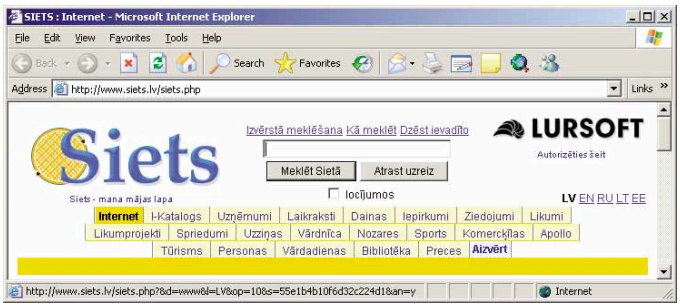
Tehnoloģijai ir divas galvenās sastāvdaļas – indeksēšanas un meklēšanas sistēma jeb tā saucamais Siets Serveris, kā arī Siets Globālais interneta rāpulis (*crawler*).

Siets Serveris ir programmatūras platforma operētājsistēmai Linux, kas darbojas kā dalītā XML datu bāze, kuras dati izvietoti uz vairākiem serveriem klastera konfigurācijā. Siets Serverī var uzglabāt jebkuras struktūras datus vairāk nekā 160 pasaules valodās, tāpēc to var uzskatīt par universālu datu uzglabāšanas un meklēšanas platformu. Siets Serveris automātiski izveido pilna teksta indeksu jebkuriem tajā ievadītajiem datiem. Pilna teksta indeksa izmēri ir salīdzināmi ar pašu oriģināldatu apjomu. Šī ir būtiskākā šādu tehnoloģiju atšķirība no SQL risinājumiem, kur indeksēti tiek tikai atsevišķi datu lauciņi un indekss ir neliels, salīdzinot ar pašu datu apjomu.

Indeksa uzglabāšanai un meklēšanai meklētājservisa Siets.lv vajadzībām šobrīd tiek izmantoti pieci Siets klastera serveri. Tādā veidā tiek atrisināta datu uzglabāšanas un ātrdarbības problēma Latvijas interneta satura indeksam. Kopējā indeksa sastāvdaļas, kas sadalītas starp serveriem, ir relatīvi nelielas, tāpēc visas indeksēšanas un meklēšanas operācijas paralēli izmanto visu serveru operatīvo atmiņu un disku sistēmas. Palielinoties indeksējamo lapu skaitam Latvijas internetā, šādu sistēmu var mērogot ar elementāru paņēmienu – pievienojot papildu serverus šim klasterim. Arī atsevišķa servera tehniskie parametri šādā sistēmā var būt salīdzinoši pieticīgi, taupot izmaksas uz aparatūru. Visbeidzot, šādā sistēmā reti kurš no lietotājiem pamana aparatūras tehniskās kļūmes, jo, sabojājoties vienam no klastera serveriem, meklēšanas pakalpojumu serviss turpina darboties. Pēc analoģiska principa – kā daudzu Linux datoru klastera risinājums – darbojas pasaulē populārāko meklētāju Yahoo un Google *virtuves* tehnoloģijas.



1. attēls. Detalizēts tehnoloģijas apraksts atrodams www.siets.net.



2. attēls. SIETS.LV piedāvā arī iespēju meklēt vārdus latviešu valodas locījumos.

Otra svarīgākā sastāvdaļa tehnoloģijā Siets ir Siets interneta rāpulis. Vienkāršotā variantā šī programmatūra ir iekļauta Siets Servera komplektā, kuru lietotāji var bez maksas lejupielādēt un bez maksas izmantot, lai indeksētu līdz 20 000 dokumentiem. Šī datu savākšanas programmatūra noder, lai bez lielas programmēšanas varētu izveidot no saviem datiem tikpat ātru un spēcīgu meklēšanas funkciju, kāda jau darbojas Siets.lv interneta meklētājā. Pietiek ievadīt atbilstošo *Web* vai intraneta serveru resursu nosaukumus, un šo serveru saturs tiks savākts, nokonvertēts uz tekstuālo informāciju un noindeksēts Siets datubāzē. Siets rāpulis atbalsta ne tikai *Web* lapu, bet arī Word, Excel, PDF, PowerPoint, RTF, Postscript un citu failu formātu dokumentu indeksēšanu.

Siets Globālā interneta rāpuļa programmatūra atšķiras no vienkāršotās Siets rāpuļa versijas ar to, ka automātiski seko līdzi interneta saitēm un pakāpeniski apstaigā visus kādas valsts interneta domēnus, kas atbilsts kādam norādītam augšējam domēnu līmenim, piemēram, Latvijas gadījumā tas ir .LV domēns. Šādas sistēmas savāktais datu daudzums ir ļoti liels, tādēļ Siets Globālais interneta rāpulis darbojas klastera konfigurācijā.

KĀDAS IR PRIEKŠROCĪBAS LATVIJAS INTERNETA MEKLĒTĀJIEM

Daudzi lietotāji, kas līdz šim ir lietojuši Google meklētāja latviskoto versiju, šodien jau ir atklājuši, ka labākus meklēšanas rezultātus par Latvijai specifisku saturu bieži vien var iegūt, izmantojot vietējos interneta meklētājus.

Vietējiem interneta meklētājiem ir virkne priekšrocību, no kuriem galvenās ir lielāks vietējo datu apjoms un biežāka to indeksa aktualizācija. Gan Google, gan Yahoo grēko ar to, ka šo meklētāju rāpuli visbiežāk pārstaigā Latvijas serveru dažas galvenās lapas, taču lielāko daļu to pārējā satura mēdz aktualizēt samērā reti – vidēji reizi mēnesī. Bieži vien pasaules meklētājos daudzās lapās sameklējamā informācija par Latvijas internetu jau ir novecojusi. Vietējās interneta meklēšanas sistēmas aktualizē datus daudz biežāk, piemēram, meklētājs Siets.lv pilnībā atjauno visu Latvijas interneta indeksu reizi nedēļā. Vēl viena priekšrocība – vietējie meklētāji parasti indeksē 2–4 reizes vairāk lapu no Latvijas interneta satura nekā pasaules vadošie meklētāji. Pasaules meklētāju indeksējamo lapu skaitu arī ierobežo brīžiem sliktie interneta sakari ar Latvijas interneta serveriem.

Lietotāju atzinību ir ieguvusi arī dažu vietējo interneta meklētāju piedāvātā iespēja meklēt vārdus latviešu valodas locījumos, ko nepiedāvā neviens no vadošajiem pasaules meklētājiem. Latvijā raksta tapšanas brīdī šāda iespēja bija SIETS.LV (sk. 2. attēlu) un LETONIKA meklētājiem.

Daudzi diskutē par meklēšanas rezultātu kvalitāti, taču var pārliecināties, ka, piemēram, gan Google, gan Siets meklētājos tā ir ļoti līdzīga. Visi labākie interneta meklētāji nosaka lapu svarīgumu pēc tā, cik daudz citās interneta lapās iekļautas saites uz konkrēto lapu. Tādā veidā tiek noteikta konkrētu lapu atrašanās vieta meklēšanas rezultātos, par kuru faktiski balso visa interneta sabiedrība. Jo vairāk meklējamajam saturam atbilstošu saišu ir citās *Web* lapās, kas visas norāda uz kādu konkrētu Web lapu, jo vairāk "balsu" šī lapa saņem interneta meklēšanas rezultātos. Lietotājiem zaudējot interesi par kādu resursu, atsauksmju skaits samazinās un lapas reitings arī ar laiku sarūk. Šādā veidā internetā notiek sava veida pašregulācija, kas garantē pietiekami augstu meklēšanas rezultātu kvalitāti un aktualitāti lietotāju acīs.

KĀDUS PIELIETOJUMUS VAR VEIDOT AR INTERNETA MEKLĒTĀJU TEHNOLOĢIJĀM

Izmantojot to pašu tehnoloģisko risinājumu, kas noder interneta meklētāju izveidei, iespējams realizēt visdažādākos tā saucamos vertikālos meklētājus par konkrētu industriju, interešu sfēru vai objektu grupu. Tādā veidā realizēts, piemēram, Google attēlu meklētājs, vai Siets.lv Latvijas meklētājservisā – meklēšana vairāk nekā 20 citos resursos tādos kā Uzņēmumi, Laikraksti, Personas, Spriedumi, Likumi, Nozares, Dainas, Katalogs u.c.

Tā kā visa publicētā informācija internetā faktiski satur hipersaites jeb tekstuālas norādes, ar šo tehnoloģiju ir iespējams savākt un apkopot datus pilnteksta meklēšanai par jebkura formāta informāciju – attēliem, mūziku, biroja dokumentiem, e-pasta ziņojumiem, *Web* portālu komentāriem u.tml. Ar īpašu datu bāzu pieslēguma utilītu iespējams apkopot arī datus pilnteksta meklēšanai no uzņēmumu SQL datu bāzēm.

Ja ir interese veidot meklētājus par atsevišķu sfēru internetā, tad iespējams, izmantojot dažādus katalogu servisus, indeksēt tikai konkrētas sfēras uzņēmumu vai organizāciju *Web* resursus un izveidot šādā veidā labus meklētājus par tūrismu, grāmatām, sportu u.tml.

Interesanti ir tas, ka šobrīd visstraujāko izaugsmi pasaulē prognozē tieši vertikālu interneta meklētāju tirgum. Iemesls ir vienkāršs – vispārīgajos interneta meklētājos pasaulē (tādos kā Google) ir indeksēts tikai nepilns 1% no globālā tīmekļa satura (daži avoti nosauc skaitli 2%, bet tas lietas būtību nemaina). Latvijas populārākie interneta meklētāji indeksē vidēji 2–4 reizes vairāk Latvijas datu nekā pasaules vadošie meklētāji, taču arī tas sastāda tikai 2–4% no Latvijas internetā publicētā satura.

Tā kā vismaz 96% no iterneta pieejamās informācijas joprojām nav ātri atrodama nevienā interneta meklētājā, tad izaugsmes iespējas šajā tirgū ir acīmredzamas.

Latvijas uzņēmumiem ir iespējams ar dažādiem inovatīviem risinājumiem apsteigt pasaules līderus. Piemēram, Siets Servera programmatūras platformā ir iebūvēta iespēja veidot dažādus vietējo resursu meklētājus, izmantojot geodatus, kur meklējamais saturs tiks parādīts sakārtojumā pēc tuvākās distances no konkrētas ģeogrāfiskās koordinātas – garuma un platuma. Piemēram, Siets tehnoloģija ir piegādāta ASV biznesa direktoriju interneta servisam MyPages *(www.mypages.com/home.html),* kas ar tās palīdzību katru dienu apstrādā vairāk nekā 700 000 meklēšanas pieprasījumus par 18 miljoniem ASV uzņēmumu adrešu. Līdzīgas meklēšanas sistēmas, kas "jūt lietotāja atrašanās vietu", iespējams veidot dažādiem GPS un mobilajiem lietojumiem.

Gints Ernestsons,
Lursoft tehniskais direktors,
gints@lursoft.lv