KLAIPĖDOS UNIVERSITETAS
GAMTOS IR MATEMATIKOS MOKSLŲ FAKULTETAS
INFORMATIKOS KATEDRA

ALEKSANDR SMIRNOV
Informatikos studijų studentas

TINKLAPIŲ OPTIMIZAVIMO PAIEŠKOS SISTEMOMS TYRIMAI

Baigiamasis magistro darbas

Mokslinis vadovas
doc. dr. Vitalijus Denisovas

KLAIPĖDA, 2005

Klaipėdos universitetas
Gamtos ir matematikos mokslų fakultetas
Informatikos katedra

Baigiamasis magistro darbas

TINKLAPIŲ OPTIMIZAVIMO PAIEŠKOS SISTEMOMS TYRIMAI

Aleksandr Smirnov

ANOTACIJA

Vis sparčiau į mūsų gyvenimą žengiantis pasaulinis tinklas Internet įtakoja naujas verslo sritis. Sąvoka „tinklalapis“ darosi žinoma ir vis svarbesnė verslininkams, o reklama Internete vis populiaresnė, nes komercinių tinklapių lankytojai – tai potencialūs įmonių ir įstaigų klientai. Pastebėta, kad internetiniai projektai dažnai žlunga – tinklapis gražus, funkcionalus, jame daug naudingos informacijos, bet tai neduoda beveik jokio naudingo rezultato arba rezultatas toli gražu neatspindi tikslų. Taip atsitinka dėl vienos paprastos priežasties: apie šį tinklapį tiesiog niekas nežino!

Tarp tūkstančių tinklapių, rastų su paieškos sistema, vartotojas gali peržiūrėti ir įvertinti tik keletą paieškos rezultatų. Anksčiau ar vėliau tinklapių savininkai tai pastebi, ir tada jiems kyla klausimas: „Kodėl pirmajame paieškos sistemos rezultatų puslapyje yra ne mano tinklapis?“ Tuo labiau, kad jo tematika visiškai atitinka užklausą ir turinio kokybė ne blogesnė, o dažnai net geresnė už konkurentų resursus. Atsakymas yra toks: taip yra todėl, kad tinklapis nėra optimizuotas – t.y. nėra pritaikytas paieškos sistemų atrankos ir reitingavimo algoritmams.

Pagrindinis šio baigiamojo magistro darbo tikslas yra išanalizuoti, atrinkti ir praktiškai pritaikyti efektyvias tinklapių optimizavimo paieškos sistemoms strategijas ir metodus.

Darbe išsamiai išanalizuota pasaulinio tinklo informacijos dinamika ir jos struktūra, ištirti šiuo metu populiariausios paieškos sistemos Google veikimo principai ir algoritmai bei kitų paieškos sistemų ir katalogų funkcionavimo ypatumai, išanalizuoti tinklapių reitingo paieškos sistemose sudarymo algoritmai. Darbe pasiūlytos strategijos ir metodai yra praktiškai įgyvendinti kompanijos ProfIS vykdomuose projektuose.

Klaipėda University
Faculty of Natural Sciences and Mathematics
Computer Science Department

Master thesis

RESEARCHES IN OPTIMISATION OF WEBSITES FOR THE SEARCH ENGINES

Smirnov Aleksandr

ANNOTATION

World Wide Web affects the new areas of business as it occurs in our lives more and more often. The term “site” becomes well-known and important for the employers and the Internet advertisement more popular since the commercial sites’ visitors are potential clients of the enterprises and institutes. It is noticed that Internet projects often fail: despite of the site is fine, functional and has a lot of useful information, the result is useless or it does not corresponds to its purposes. This happens because of the very simple reason: nobody knows about the site!

Among thousands of sites which can be found by the search engine the user can overview and evaluate only few of the search results. Sooner or later sites’ owners find out this fact, and they question themselves: “Why not my site is placed in the first search engine result page?” Moreover, the site’s theme completely corresponds to the query, and the quality of content is fine and even better than the one of competitors. The answer exists: the site is not optimized for the search engines.

The main purpose of the paper is to analyze, select and apply in practice the effective search engine optimization strategies and methods.

The main contribution of the paper is the detail description of information dynamics and its structure, research of the world popular search engine’s Google work principles and also other search engines’ and directories’ functional peculiarities, analysis of the page ranging algorithms. The methods and strategies offered in this work are practically implemented in the projects of ProfIS Company.

PADĖKA

Studijos Klaipėdos universitete padėjo man surasti savo vietą gyvenime, suprasti, kiek daug mūsų dienų pasaulyje reiškia žinios ir kaip svarbu būti tarp žmonių, kurie savo žinias noriai perteikia kitiems.

Nuoširdžiai dėkoju gerb. dėstytojui docentui Vitalijui Denisovui už pagalbą rašant magistro darbą, naudingus patarimus ir vadovavimą visų studijų metu.

Dėkoju visam Informatikos katedros dėstytojų kolektyvui už žinias, įgytas per visus studijų metus, ir patirtus įspūdžius.

Aleksandr Smirnov

TERMINŲ IR SANTRUMPŲ ŽODYNAS

Baneris	Reklaminis skelbimas paveikslėlio pavidalu.
DB (Data Base)	Duomenų bazė.
Lematizavimas	Žodžių sutvarkymas pagal kanoninę formą.
LPM	Loginės paieškos modelis.
Pertinentiškumas	Naudingos informacijos apimties santykis su visa gautos informacijos apimtimi.
PR (PageRank)	Algoritmo „PageRank“ sutrumpintas pavadinimas.
Relevantiškumas	Formalus informacijos, kurią pateikia paieškos sistema, atitikimas vartotojo užklausai.
SCC (Strongly Connected Component)	Stipraus ryšingumo komponentai
SE (Search Engine)	Paieškos sistema.
SEO (Search Engine Optimization)	Tinklapių optimizacija paieškos sistemoms.
SERP (Search engine result page)	Paieškos sistemos rezultatų išvedimo puslapis
Spamas	Beprasmingi paieškos rezultatai arba reklama.
VEM	Vektorinės erdvės modelis

TURINYS

ĮVADAS

Darbo apžvalga

Vis sparčiau į mūsų gyvenimą žengiantis pasaulinis tinklas Internet įtakoja naujas verslo sritis. Sąvoka „tinklalapis“ darosi žinoma ir vis svarbesnė verslininkams, o reklama Internete vis populiaresnė, nes komercinių tinklapių lankytojai – tai potencialūs įmonių ir įstaigų klientai. Dar visai neseniai svarbiausias dėmesys buvo skiriamas tik įmonės tinklapio dizainui, - buvo manoma, kad to užtenka, tačiau vėliau pastebėta, kad internetiniai projektai labai dažnai nepasiseka – tinklapis gražus, funkcionalus, jame daug naudingos informacijos, bet tai neduoda beveik jokio naudingo rezultato arba rezultatas toli gražu neatspindi tikslų. Tai atsitinka dėl vienos paprastos priežasties: apie šį tinklapį tiesiog niekas nežino!

Interneto reklamos rinka jau sėkmingai konkuruoja su tradicinės reklamos rinka. Daugeliui vidutinio ir smulkiojo verslo atstovų kiekis klientų, atėjusių iš paieškos sistemų dėl tinklapių optimizacijos, dažnai ir viršija kiekį klientų, atėjusių dėl kitų internetinės reklamos rūšių. Dažnai kokybinis tinklapių turinys ir jo optimizacija, atitinkanti paieškos sistemų reikalavimus, leidžia žymiai sumažinti išlaidas, kurių paprastai reikalauja brangi neinternetinė reklama, o kartais ir išvis jos atsisakyti.

Svarbu žinoti, kad tarp tūkstančių tinklapių, rastų pagal paieškos užklausą, vartotojas gali peržiūrėti ir įvertinti tik keletą pateikiamų pirmame ir antrame paieškos rezultatų puslapyje. Kaip tik komercinio tinklapio buvimas pirmame rezultatų dešimtuke pagal tikslinę užklausą gali garantuoti maksimalų ypač suinteresuotų pateiktų paslaugų ar prekių vartotojų kiekį. Tuo tarpu tokio vartotojo pritraukimas žymiai pigesnis, negu atvykusio į tinklapį pagal reklamą.

Pagal paieškos sistemos AllTheWeb (www.alltheweb.com) [22] ir analitinės kompanijos NPD Group (www.npd.com) duomenis ir pagal asmeniškus tyrimus 2005 m. - 85% tinklapio lankytojų ateina iš paieškos sistemų. Ir 9 iš 10 atėjusių į tinklapį per paieškos sistemą – tiksliniai lankytojai. Lankytojas dažniausiai peržiūri tik pirmąsias 3 paieškos rezultatų nuorodas. 70% lankytojų peržiūri 10 pirmųjų nuorodų, ir tik 5% gilinasi į antrą rezultatų puslapį.

Anksčiau ar vėliau tinklapių savininkai tai pastebi, ir tada jiems kyla klausimas: „Kodėl pirmajame paieškos rezultatų puslapyje yra ne mano tinklapis?“ – tuo labiau, kad jo tematika visiškai atitinka užklausą ir turinio kokybė ne blogesnė, o dažnai net geresnė už konkurentų resursus. Dažnai atsitinka taip, kad pagal garsaus prekinio ženklo užklausą ženklo savininko tinklapis neužima reikšmingos pozicijos.

Atrodytų, kad vienintelė išeitis iš susiklosčiusios padėties – brangūs indėliai į mokamas paieškos sistemas ir raktinių žodžių pirkimas, tačiau kartu egzistuoja viena svarbi problema: vartotojai nemėgsta reklamos. Absoliuti jų dauguma nekreipia dėmesio į mokamus skelbimus,

teisingai atiduodami pirmenybę „natūraliems“ paieškos rezultatams. Tarp kitko, pirmųjų vietų mokamų paieškos sistemų reitinguose netenkama, kai tik nustojama jas finansuoti. Ar egzistuoja efektyvesnė, ilgalaikė ir nebrangi alternatyva? Taip. Tai vadinamoji „paieškos optimizacija“ (toliau SEO nuo angl. «Search Engine Optimization»).

Kaip tik šita alternatyva pagimdė naują profesiją, kurios atstovai vis reikalingesni paieškos paslaugų rinkoje. Tai tinklapių pagal paieškos sistemų ypatumus optimizatoriai.

Dar visai neseniai optimizacijos rinka tik pradėjo egzistuoti, bet šiuo metu ji paklausi visame pasaulyje. Lietuvoje tuo užsiima tik maža dalis kompanijų, kurių profilis – tinklapių kūrimas, tačiau visos vystymosi prielaidos egzistuoja.

Darbo tyrimo objektai:

- Pasaulinio tinklo (Internet) informacijos dinamika.
- Paieškos sistemos (search engines).
- Tinklapiai ir svetainės (websites).

Darbo tikslas:

- Išanalizuoti, atrinkti ir praktiškai pritaikyti efektyvias tinklapių optimizavimo paieškos sistemoms strategijas ir metodus.

Darbo uždaviniai:

- Išanalizuoti pasaulinio tinklo informacijos dinamiką ir jos struktūrą.
- Ištirti šiuo metu populiariausios paieškos sistemos Google veikimo principus ir algoritmus bei kitų paieškos sistemų ir katalogų funkcionavimo ypatumus.
- Išanalizuoti tinklapių reitingo paieškos sistemose sudarymo algoritmus.
- Pasiūlyti ir įgyvendinti praktiškai naudingus paieškos optimizavimo metodus.

Temos naujumas ir aktualumas

Tema yra labai aktuali. Pirmieji straipsniai apie paieškos sistemas pasirodė tik apie 1999 m. Šitoje srityje iki šiol egzistuoja nemaža neištyrinėtų temų – pvz. tinklapių optimizacija, kuriai ir skiriamas šis darbas. Atsižvelgiant į paieškos sistemų vystymąsi ir didėjantį komercinių struktūrų dėmesį joms, optimizavimo tematika šiuo metu darosi labai aktuali, tačiau mokslinių ir metodinių straipsnių šia tema vis dar yra labai mažai. Efektyvus SEO panaudojimas kuriant tinklapius leidžia žymiai padidinti tinklapių lankomumą (tai reiškia ir verslo pelningumą). Kadangi paieškos sistemos pagal savo prigimtį nuolatos kinta, tyrimai šioje srityje taip pat turi būti vykdomi nuolat ir sistemingai.

Tyrimo metodika

Tyrimai atlikti naudojant literatūros apžvalgos, lyginamosios analizės, sisteminės analizės ir kompiuterinių eksperimentų metodus.

Mokslinė darbo vertė

Darbe pasiūlyti metodai ir tyrimai leidžia sėkmingai taikyti paieškos optimizaciją kuriant įvairaus sudėtingumo lygio tinklapius ir svetaines.

Darbo rezultatai

- Atlikti paieškos sistemų tyrimai ir paskelbti pasiekti rezultatai.
- Eksperimentais su veikiančiais įvairių įmonių tinklapiais parodytas paieškos optimizacijos (SEO) svarbumas.
- Pasiūlyti ir aprašyti efektyvūs SEO metodai ir strategijos.
- Pateikta išsami informaciją apie paieškos sistemas ir paieškos robotus.
- Pasiūlyti naudingi praktiniai patarimai tinklapiams pagal paieškos sistemų reikalavimus optimizuoti.

Darbo praktinis panaudojimas

Autoriaus pasiūlyti tyrimai, metodai ir strategijos praktiškai įgyvendinti kompanijos “ProfIS” (www.profis.lt) vykdomuose projektuose.

Darbo sandara

Vykdant kokybišką tinklapio optimizaciją, reikia suvokti Interneto informacijos struktūrą ir paieškos sistemų bei katalogų veikimo principus, - tai leis sukurti optimalaus jų galimybių panaudojimo strategiją. Dėl tos priežasties darbe išskirtos kelios dalys:

- Pirmoje dalyje aprašyta pasaulinio tinklo istorija ir jo informacijos dinamika.
- Antroje darbo dalyje išanalizuotas paieškos sistemų darbas remiantis šiuo metu populiariausia paieškos sistema „Google“ (www.google.com).
- Trečioje dalyje aprašyti tinklapio reitingo apskaičiavimo algoritmai, remiantis žinomo PageRank algoritmo pavyzdžiu.
- Ketvirtoje dalyje aprašyti paieškos optimizacijos tyrimai, metodai ir strategijos bei pateikti praktiniai sprendimai ir rekomendacijos tinklapių kūrėjams.
- Penktoje dalyje trumpai peržiūrėtos paieškos sistemų, internetinės reklamos ir paieškos optimizacijos rinkos vystymosi perspektyvos.

1. PASAULINIO TINKLO DINAMIKA

1.1. Žmogus ir informacija

Informacijos kiekis, tenkantis žmogui dabarties pasaulyje, lemia svarbių duomenų atskyrimo nuo informacinio triukšmo aktualumą. Žmogus, žmonių grupė, informacinė tarnyba, profesionalūs ekspertai-analitikai jau nebegali apdoroti šiuolaikinių informacinių srautų. Dažnai net patyrę ekspertai neranda žinių, kad galėtų priimti sprendimus, ir todėl atskirų žmonių, kolektyvų, net valstybių veiksmai neatitinka realios padėties. Todėl pati svarbiausia šiuolaikinių komunikacijų problema – rasti informacijos jūroje iš tiesų svarbią informaciją [22].

Informacijos gausa jau seniai suvokiama kaip savaime suprantamas dalykas. Kiekybinis jos apimties vertinimas kažin ar gali tapti ypatingų apmąstymų priežastimi. Tačiau jeigu tokius požymius struktūriškai išanalizuotume, tai gauti rezultatai gali būti labai nelaukti.

Pavyzdys – informacijos kiekio pasaulyje per metus tyrimai. Nuo 2000 m. jie vykdomi Berklio universitete Kalifornijoje vadovaujant profesoriams Piteriui Laimanui (Peter Layman) ir Holui Varienui (Hal Varien) – Projektas „How Much Information“, (http://www.sims.berkeley.edu/research/projects/how-much-info). Mokslininkai padarė išvadą, kad per trejus metus iki 2002 m. informacijos, sukurtos žmonijos, kiekis padvigubėjo! O 2002 m. buvo sukurta 5 milijonai terabaitų informacijos. Palyginkime: JAV Kongreso bibliotekos fondo, kuriame yra 19 mln. knygų ir 56 mln. rankraščių, apimtis - apie 10 terabaitų. Šitame tyrime informacija buvo struktūrizuota pagal laikmenų tipus. Pasirodė, kad vyrauja magnetinės laikmenos – jų yra 90%. Dauguma jų yra kietieji diskai. Kino ir foto juostos, spaudiniai ir kiti popieriniai dokumentai bei optinės skaitmeninės laikmenos sudaro tik 7% informacijos.

1.2. Interneto paradoksas

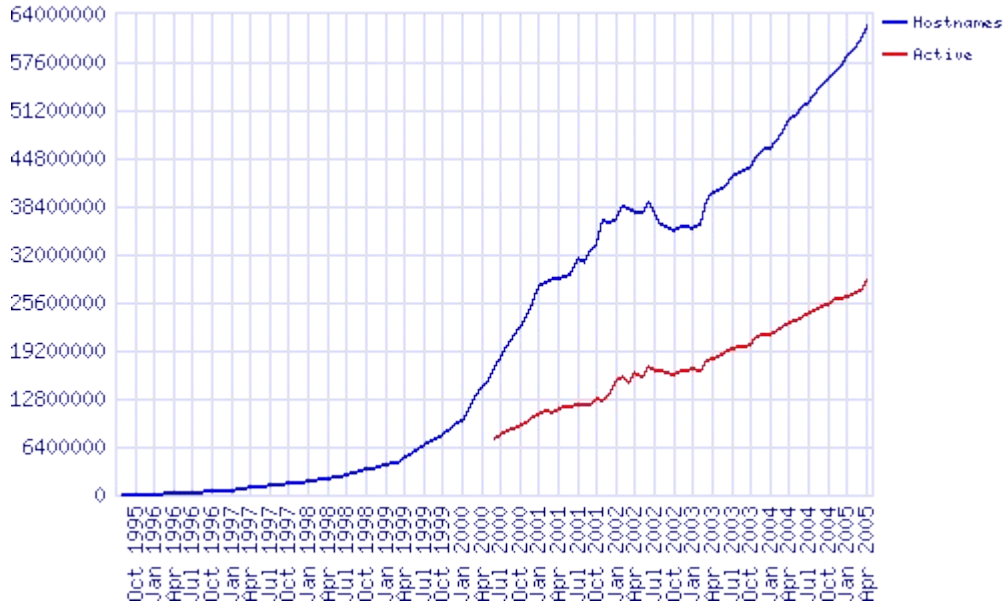
Šiandien kiekvienas Interneto vartotojas pats pajunta vieną didžiausių šios aplinkos paradoksų: **naudingos informacijos daugėja, bet surasti ją tampa vis sunkiau.** Internetas per daugiau kaip 30 savo gyvavimo metų peržengė karinių JAV laboratorijų (kur pasaulinis tinklas gimė kaip projektas ARPANET 20 a. 60-aisiais metais) ir mokslinių tyrimų ribas ir šiuo metų tapo vienu žinomiausiu pasaulinių reiškinių [31]. Kodėl įvyko taip, kad iš šimtų kompiuterinių tinklų taip išsivystė tik Internetas? Atsakymų keletas:

- Aukštas technologiškumas, patikimumas ir galimybė veikti ekstremaliausiomis sąlygomis.

- Protokolų atvirumas ir prieinamumas kiekvienam. To rezultatas – tinklą palaikė dauguma vartotojų ir stambiausių programinės įrangos ir aparatūros gamintojų.
- Sistemos sugebėjimas savarankiškai vystytis ir plėstis.

Tai paaiškinama tuo, kad kuo daugiau resursų įtraukiama į tinklą, tuo jis tampa įdomesnis ir naudingesnis vartotojams, kurių kiekis atitinkamai auga. Yra ir kita priežastis: nuolatinis išlaidų darbui Internete mažėjimas. Pagal analitinės kompanijos **Cyveillance** (http://www.cyveillance.com) pranešimą, tinklo augimo tempai sudaro 7 mln. naujų puslapių per metus. Augimo centras dabartiniu metu traukiasi iš JAV. Informacinių resursų tinkle apimties augimo dinamika tokia didžiulė (galima pažymėti, kad 1998 m. tinklapių kiekis pasaulyje sudarė tik 1 mln.), kad informacijos Internete apdorojimo metodai ne visiškai akivaizdūs.

Kaip skelbiama publikuotoje analitinės tarnybos Netcraft Web Server Survey [26] ataskaitoje (1 pav.), tinklapių kiekis Internete 2005 m. balandžio mėnesį pasiekė **62,286,451** (iš jų daugiau kaip pusė aktyvių).



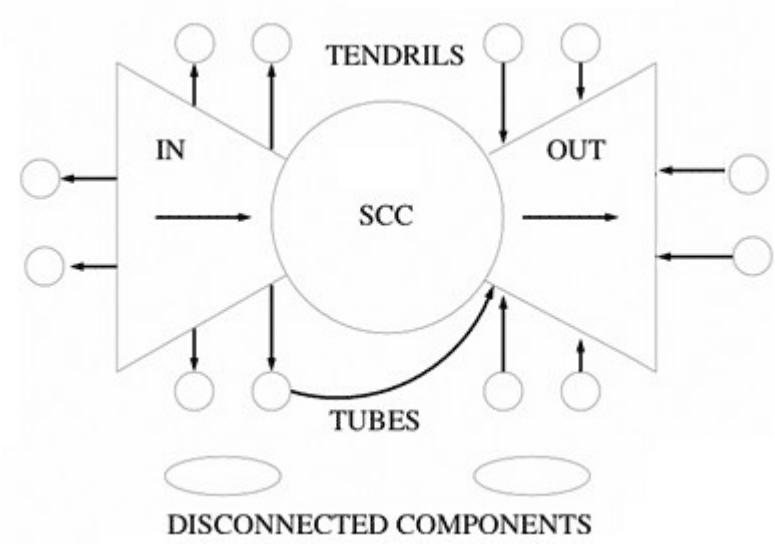
1 pav. ***Tinklapių kiekis pasaulyje 1995.08 – 2005.04 m.***

Pagal informacijos atnaujinimą visa Interneto erdvę galima padalinti į dvi sudedamąsias dalis: stabilią ir dinamiškąją. Stabilioje dalyje yra ilgalaikė informacija, pvz. monografijos, galerijos, kolekcijos ar archyvai. Dinamiškojoje dalyje yra nuolatos atnaujinami arba nauji resursai. Nedidelė jos dalis paskui įsilieja į stabiliąją, o tuo tarpu didžioji dalis dingsta iš tinklo.

1.3. Web-erdvės topologija

Skirtingai nuo įprastos informacijos saugyklos, web-erdvė turi didelį kiekį joje paslėptų neakivaizdžių ekspertinių vertinimų, realizuotų kaip hypernuorodos. Būtent hypernuorodos tapo web-erdvės modelio kūrimo baze. Norint paieškos sistemų priemonėmis apimti kuo daugiau informacinių resursų, būtina atsižvelgti į visos web-erdvės architektūrą, bet kaip tik šios informacijos anksčiau niekas neturėjo. Iki 1999 m. artimas realybei matematinis modelis neegzistavo [11].

1999 m. lapkritį Andrei Broderis (Andrei Broder), vienas IBM tekstų paieškos ir analizės Instituto vadovų, ir jo bendraautoriai iš AltaVista ir Compaq kompanijų įvykdė šuolį, matematiškai aprašė resursų ir egzistuojantį web-erdvės ryšių žemėlapį [7]. Tyrimai paneigė paplitusią nuomonę kad Internetas - tai vieninga tanki erdvė. Peržiūrėję su paieškos sistemos AltaVista pagalba daugiau kaip 200 mln. tinklapių ir keletą milijardų nuorodų, patalpintų tuose tinklapiuose, mokslininkai padarė konkrečias išvados apie web-erdvės struktūrą [6]. Jų nuomone, ši struktūra realybėje atitinka dinamiškai orientuotą grafą su „kaklaraiščio-peteliškės“ topologija (Bow Tie) (2 pav.), kurio viršūnės atitinka tinklapius, o briaunos – tinklapius jungiančias hypernuorodas [2].



2 pav. ***„Kaklaraiščio-peteliškės“ topologija (http://wwwmayr.informatik.tu-muenchen.de/spp1126/stand.html)***

Atskirų tinklapių ryšių struktūros analizė, atlikta šio modelio rėmuose, leidžia nustatyti tokius faktus:

- **Centrinis branduolys** (26% tinklapių) — stipraus ryšingumo komponentai (angl. „SCC“, Strongly Connected Component), arba kaklaraiščio mazgas. Jam priklauso tinklapiai, surišti taip stipriai, kad paprasčiausiai sekant hypernuorodas iš bet kurios jų galima patekti į bet kurią kitą.
- **Išvykimo** (IN, Upstream) (22% tinklapių) — jie turi hypernuorodas, kurios galiausiai veda į branduolį, bet iš branduolio į jas patekti negalima.
- **Atvykimo** (OUT, Downstream) (22% tinklapių) – į juos galima patekti pagal hypernuorodas, bet negalima grįžti į branduolį.
- **Ataugos** (Tendrils) (22% tinklapių) – tai visiškai izoliuoti nuo centrinio branduolio tinklapiai. Tai arba „**ragai**“, surišti nuorodomis su bet kurios kitos kategorijos tinklapiais, arba „**sąsmauka**“ (Tubes), jungianti du tinklapius, neįeinančius į branduolį.
- **Salos** (Isles or Disconected Components) – tinklapiai, kurie visiškai nesusikerta su kitais Interneto resursais. Vienintelė galimybė surasti šitos grupės resursus – žinoti jų adresus. Jokios paieškos sistemos nesugebės rasti šitų salų, jei anksčiau jos nebuvo kokiu nors būdu sujungtos su kitomis Interneto dalimis, todėl apskaičiuoti šių tinklapių procentinę dalį šioje schemoje neįmanoma.

Mokslininkai pastebėjo, kad šitų keturių kategorijų proporcijos beveik nesikeičia, nors bendra Interneto resursų apimtis auga. Buvo ištirti tokie pateikto modelio parametrai, kaip vidutinis tinklapių, per kuriuos jungiasi hypernuorodomis du tinklapiai, kiekis, ir įeinančių bei išeinančių hypernuorodų pasiskirstymas [19].

Buvo įrodyta, kad web-erdvės grafo viršūnių išėjimo ir įėjimo puslaipsnių pasiskirstymas priklauso nuo laipsnio dėsnio („power law“), t.y tikimybė, kad atitinkamas viršūnės laipsnis lygus **„n“** – proporcingas n^{-k} (pagal Kalifornijos universiteto skaičiavimus, įeinančioms nuorodoms $k \approx 2,1$, o išeinančioms $k \approx 2,45$) [33]. Dviejų tinklapių ryšio tikimybė – **24%**. Vidutinis tinklapių, per kuriuos jungiasi hypernuorodomis du tinklapiai, kiekis yra **16-20.** Jeigu šitas kelias dvipusis, tai tas kiekis sumažėja iki **7.**

Informacija apie web-erdvės struktūrą jau plačiai naudojama sprendžiant daugelį uždavinių, pvz. skenavimo mechanizmų efektyvumo optimizacijai ir naujų web-servisų kūrimui, o iš dalies - paieškos sistemų kūrimui, apie kurias toliau bus kalbama.

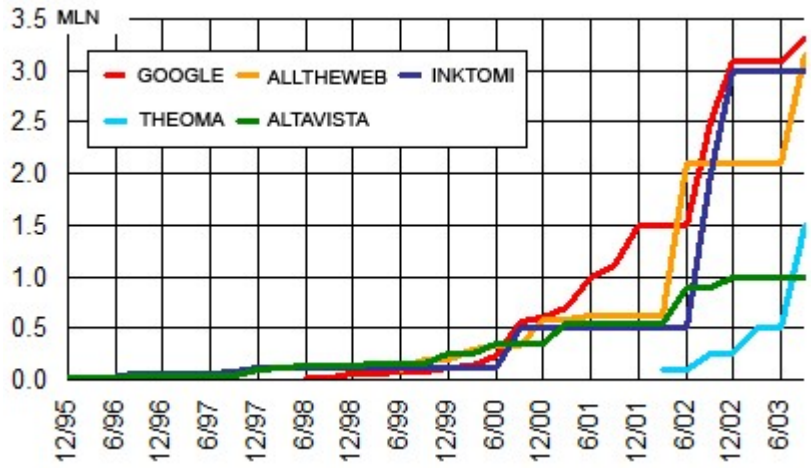
2. PAIEŠKOS SISTEMOS

2.1. Internetinių paieškos sistemų vystymosi istorija

Čia ir toliau terminu „paieškos sistema“ bus aprašoma internetinė paieškos sistema. Kad būtų įmanoma rasi reikalingą informaciją Internete – stambiausioje pilnatekstėje duomenų bazėje, būtina naudotis labai galingomis paieškos sistemomis, kurios jau dešimtmetį vystosi ir informacinių technologijų rinkoje konkuruoja viena su kita.

Skaičiavimo technikos ir kompiuterinių ryšių raida nulėmė pasirodymą sistemų, kurių tikslas – paieška pilnatekstinių dokumentų masyvuose. Tokiems dokumentams priklauso straipsniai, norminiai aktai, referatai, brošiūros, disertacijos, monografijos. Iki tam tiko laiko pilnatekstės paieškos sistemos buvo naudojamos daugiausia specialistų, kurių skaičius nebuvo didelis: archyvų, bibliotekų darbuotojai, mokslininkai, analitikai.

Paieškos sistemų technologija turėjo labai sparčiai vystytis, kad galėtų suspėti su Interneto augimu. 1994 m. vienas pirmųjų paieškos variklių, **World Wide Web Worm (WWWW)**, turėjo surinkęs 110000 puslapių ir tinkle pasiekiamų dokumentų indeksų. 1997 m. lapkritį pagrindinės paieškos sistemos galėjo indeksuoti nuo 2 milijonų (**WebCrawler**) iki 100 milijonų internetinių dokumentų. 1994 m. kovo ir balandžio mėnesiais World Wide Web Worm apdorodavo apie 1500 užklausų per dieną. 1997 m. lapkričio mėnesį AltaVista sulaukdavo apytiksliai 20 milijonų užklausų per dieną. 2000 m. į lyderius išeina Goolge ir iki šiol ji užima pirmą vietą tarp paieškos sistemų (3 pav.) [35].



3 pav. ***Kaip keičiasi paieškos sistemų dydžiai (indeksuotų dokumentų kiekis) 1995.12 – 2003.06 laikotarpiu***

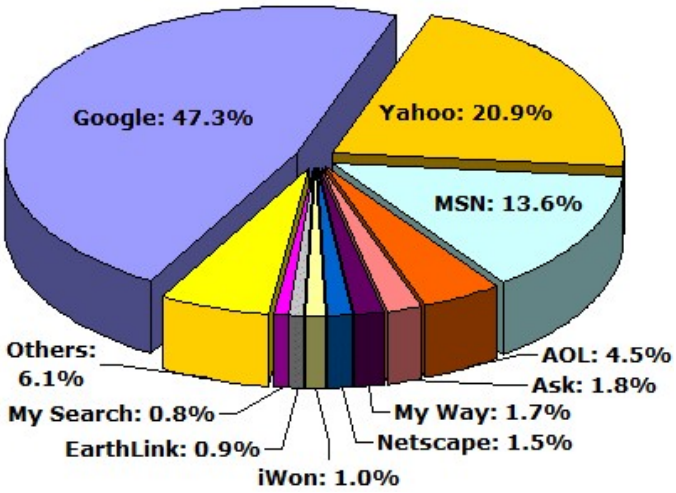
Didėjant Interneto vartotojų ir automatinių sistemų, besikreipiančių į paieškos sistemas, skaičiui, pagrindinės paieškos sistemos sulaukia milijonų užklausų per dieną. Tokių sistemų užduotis yra kokybė bei dinamiškumas. Šiuo metu išsamus indeksas gali siekti milijardus dokumentų. Tuo pat metu užklausų, kurias gauna paieškos sistema, skaičius taip pat labai sparčiai auga. Šiandien milijonams Interneto vartotojų žinomos tokios sistemos, kaip Google, Yahoo, AllTheWeb, AltaVista, kurių kiekviena apima keletą milijardų dokumentų. Mes tapome „informacinio sprogimo“ liudininkais. To sprogimo rezultatas - mažiau kaip per dešimtmetį nedaug kam žinoma pilnatekstės paieškos technologija tapo milijonams žmonių kasdieniniu įrankiu.

Kompanijos **comScore,** užsiimančios Interneto rinkos tyrimais, duomenimis, per 2004 m. balandį daugiau kaip 40 mln. pirkėjų visame pasaulyje naudojosi paieškos sistemomis (1 lentelė) [25]:

1 lentelė. ***Užklausų paieškos sistemose statistika per 2004 m. balandžio mėn.***

Šalis	Užklausų kiekis, tenkantis vienam vartotojui
Didžioji Britanija	41.8
Kanada	39.9
JAV	35.2
Prancūzija	34.0
Vokietija	31.2

Kompanijos **Nielsen NetRatings** duomeninis, 2005 m. balandžio mėnesį, paieškos sistema **Google** yra absoliutus lyderis tarp trijų pasaulyje populiariausių paieškos sistemų (4 pav.) [28]:



4 pav. ***Paieškos sistemų lankytojų statistika per 2005 m. kovo mėn.***

Įdomu, kad tos pačios kompanijos duomenimis 2004 m. gruodį **Goolge** užėmė 35% rinkos – t.y tik per pusmetį jos dalis išaugo 12%. Tokie milžiniški augimo procentai liudija, kad šiandien Google yra „paieškos sistemų karalienė“, ir kaip tik todėl šiame darbe bus detaliai analizuojamas Google veikimas.

Skirtingai nuo reliacinių DBVS, pilnatekstės paieškos sistemos neturi standartizuotos užklausų kalbos, nes kiekviena šio tipo sistema turi nuosavą paieškos kriterijų išraiškos būdą, panašų į SQL, bet turintį individualių ypatumų, susijusių su tokias momentais:

- Operacijų, priklausančių nuo žodžių tvarkos tekste, interpretacija, pvz. kontekstinio žodžių artumo operacijos.
- Rastų dokumentų atitikimas užklausoms (relevantiškumas) paieškos rezultatams pateikti.
- Nestandartinių funkcijų, pvz. dirbtinio intelekto metodų, panaudojimas (dokumentų radimas pagal panašumą, referatų kūrimas iš dokumento fragmentų).

Skirtingose paieškos sistemose skiriasi architektūra, duomenų struktūros, jų apdorojimo algoritmai, paieškos organizacijos metodologijos.

2.2. Paieškos sistemų charakteristikos

Pagrindinės paieškos sistemų charakteristikos – **paieškos apimties pilnatvė** ir **paieškos rezultatų relevantiškumas** [29]. Reikšmingos ir tokios charakteristikos, kaip užklausų apdorojimo greitis, sistemos atsiliepimo gavimas, atsakymo patikimumas ir papildomos paslaugos: galimybė surasi dokumentus, panašius į turimus (like this), galimybė prijungti automatinius vertėjus ir galimybė patikslinti užklausą tuoj po paieškos procedūros.

Paieškos apimties pilnatvė

Pirmasis paieškos apimties pilnatvės aspektas (5 pav.) – tai informacijos “aprėpimo pilnatvė**”**. Kartą sukurta Interneto resursų DB yra tinklo būsenos konkrečiu momentu lipdinys. Jeigu šita bazė nuolat ir operatyviai nebus atnaujinama, joje egzistuojančios nuorodos į dokumentus taps negyvos. Be to, DB atnaujinimo operatyvumo stoka neleis vartotojui stebėti naujausių pakeitimų jį dominančioje dalykinėje srityje. Antras aspektas – „atsiliepimo pilnatvė“ - susijęs su pilnatve informacijos, pateikiamos vartotojui pagal jo užklausą.

5 pav. ***Paieškos apimties pilnatvė***

Relevantiškumas ir pertinentiškumas

Relevantiškumas – tai formalus informacijos, kurią pateikia paieškos sistema, atitikimas vartotojo užklausai [29]. Jeigu pagal vartotojo užklausą gaunama N dokumentų, kurie yra dviejų dokumentų aibių junginys: atitinkančių užklausą (jų kiekis: N_1) ir neatitinkančių užklausos (jų kiekis: N_2), t.y. $N=N_1+N_2$. tada relevantiškumas, kaip atitikimo laipsnis, nustatomas pagal formulę:

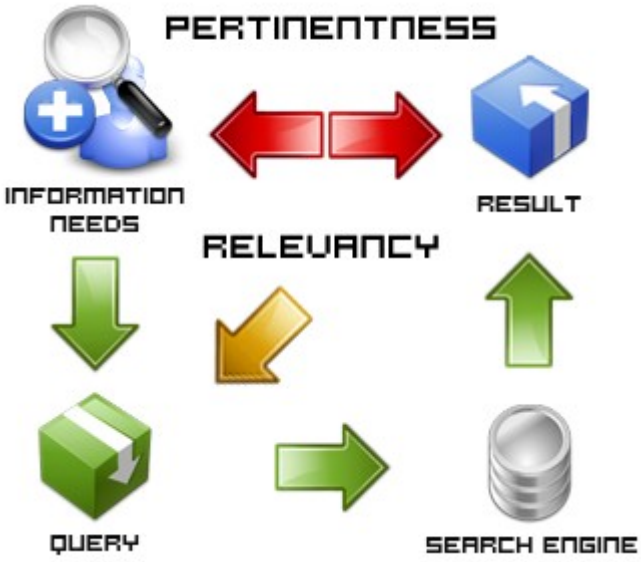
(1)

O triukšmas pagal formulę:

(2)

Jeigu pažymėsime pradiniame masyve atitinkančių užklausą dokumentų kiekį kaip R, tai

santykis nustatys paieškos apimties pilnatvę. Atitinkamai: kuo aukštesnis relevantiškumas – to žemesnė pilnatvė, ir atvirkščiai. Šitas apibrėžimas būdingas formaliam relevantiškumui, tačiau praktiškai vartojamas kitas, neformalus apibrėžimas - **pertinentiškumas** (6 pav.) – naudingos informacijos apimties santykis su visa gautos informacijos apimtimi. Vartotojui pertinentiškumas yra svarbiausias dalykas, nes formali užklausa sistemai yra kūrybiška informacinio poreikio išraiška ir ne visada atspindi patį informacinį poreikį. Daugumos vartotojų nesugebėjimas teisingai formuluoti užklausą ir gauti priimtiną atsiliepimo apimtį 20 a. pabaigoje pagimdė nuomonę, kad Internetas – milžiniškas informacijos sąvartynas.



6 pav. ***Pertinentiškumas ir relevantiškumas***

Pertinentiškumo didinimo šiuolaikinėse sistemose priemonės, be galimybės patikslinti užklausos formuluotę, turi ir svorio kategorijų, leidžiančių reitinguoti rastus dokumentus ir pateikti vartotojui svaresnius dokumentus arba riboti pateikimo kiekį. Šiuo atveju, natūralu, nukenčia pateikimo pilnumas. T.y kuo aukštesnis relevantiškumas, tuo žemesnis pilnumas, ir atvirkščiai. Relevantiškumo ir pertinentiškumo problemoms šiuolaikinėse paieškos sistemose skiriamas ypatingas dėmėsis. Taip pvz. paieškos sistema Google puikiai realizavo neformalaus relevantiškumo pasiekimo algoritmus ir dėl to tapo populiariausia paieškos sistema Internete.

2.3. Lingvistinės paieškos sistemų priemonės

Lingvistinės paieškos sistemų priemonės yra natūralios kalbos ir formalių paieškos mechanizmų sąsaja [29]. Lingvistinės paieškos sistemų priemonių elementai tai:

- Duomenų, kurie nusako architektūrą, sintaksę ir informacijos pateikimo semantiką paieškos sistemų duomenų bazėje, vaizdavimo kalbos.
- Informacinė paieškos kalba – kalba, kuria vartotojas kreipiasi į sistemą, kad gautų jį dominantį atsakymą. Šiuolaikinės informacinės paieškos kalbos turi logikos operatorius (AND, OR, NOT), kontekstinio artumo operatorių (NEAR) ir operatorių prioritetų valdymo priemones.
- Natūralios kalbos ir žymėjimo kalbos, kuriomis pateikiami dokumentai-šaltiniai.

Ypatinga reikšmė šiuolaikinėse paieškos sistemose skiriama morfologinei analizei, t.y.atskirų žodžių dokumentų tekstuose ir vartotojų užklausose apdorojimas automatiniais būdais.

2.3.1. Morfologinė analizė ir «stop-žodžiai»

Dauguma pasaulių kalbų turi pagalbinių žodžių: prielinksnių ir artikelių. Tokie žodžiai paieškos sistemose vadinami **„stop-žodžiais“** (pvz. anglų „a“, „an“, „the“, „on“ ir kiti). Paieškos sistemos paprastai į savo indeksą neįtraukia tokių žodžių, bet įskaito juos numeruodamos žodžius, o tai leidžia vykdyti paiešką frazių su stop-žodžiais, pvz. “arbata su pienu“. Stop-žodžių pašalinimas iš indekso sutrumpina indeksą ir padidina darbo efektyvumą. Tačiau kai kurios užklausos, susidedančios tik iš stop-žodžių (pvz. „to be or not to be“) negaus trokštamo atsakymo. Nepatogumų sukelia ir kai kurie polisemijos (žodžio daugiareikšmiškumo priklausymas nuo konteksto) atvejai. Pvz. angliškas žodis „**can**“ („gali“) kaip pagalbinis veiksmažodis gali būti įtrauktas į stop-žodžių sąrašą, tačiau kaip daiktavardis („konservų dėžutė“) jis yra reikšmingas.

Sukurtas sistemos žodinis indeksas daugelyje paieškos sistemų **lematizuojamas,** t.y. visi žodžiai sutvarkomi pagal kanoninę formą: daiktavardžiai – vardininko linksnis, veiksmažodžiai – bendratis ir t.t. Sistemose su morfologine analize taip pat lematizuojamos ir vartotojų užklausos. Pvz. jeigu dokumente yra derinys „baltosios naktys“, tai sistemos indekse tuo atveju bus žodžiai „baltas“ ir „naktis“. Jeigu vartotojo užklausa „balta ir naktis“ bus perduota paieškos sistemai be pakeitimų, tai reikalingas dokumentas nebūtų surastas, tačiau jeigu užklausa buvo lematizuojama, tai dokumentas bus rastas. Praktikoje lematizacija ne visada padidina pertinentiškų dokumentų kiekį. Pvz. kai kurios paieškos sistemos orientuotos tik į anglų kalbos lematizaciją ir tekstą lietuvių kalba jos priims kaip simbolių rinkinį.

2.3.2. Tezauras (Tesaurus)

Kai kurios paieškos sistemos savo darbe naudoja iš anksto paruoštus žodžių masyvus – tezaurus. Tezauruose kiekvienam leksiniam vienetui (atskiriems žodžiams, žodžių junginiams, leksiškai svarbiems sudėtingų žodžių komponentams, abreviatūroms ir sutrumpinimams) skiriamas nedidelis aiškinantis tekstas ir nuorodos į kitus šio masyvo žodžius. Nuorodos gali reikšti: sinonimiką, antonimiką, priklausomybę. Šiandien, iš vienos pusės, laikmenų tūris leidžia saugoti praktiškai neribotą tezaurų kiekį, o iš kitos pusės, tekstinės DB tokios didžiulės, kad jų indeksai dažnai viršija tezaurų tūrį, todėl geriausiai naudotis minimaliu tezauru.

2.3.3. Loginių operatorių panaudojimas

Norint įvesti sudėtingas užklausas, reikia naudotis loginiais ir kontekstiniais operatoriais, skliaustais ir t.t. Kadangi daugeliu atveju (pagal statistiką 70 % užklausų susideda iš vieno žodžio) tai nereikalinga, todėl paieškos sistemos paprastai turi dvi sąsajas: paprastą (pagal nutylėjimą) ir išplėstą. Dauguma paieškos sistemų palaiko kontekstinio artumo operacijas (NEAR, FAR) pvz. reikšmių kabutėse paieška. Pavyzdžiui, užklausa „obuoliai sniege“ ekvivalentiška užklausoms „obuoliai ir sniegas“, „obuoliai po sniegu“ ir „obuolys sniegas“.

2.3.4. Paieškų tipai

Paieška pagal parametrus

Dokumentų paieškos pagal parametrus galimybė leidžia apriboti paieškos diapazoną iki paieškos pagal URL reikšmes, pagal datas, pavadinimus ir t.t. Dažniausiai rasti paieškos pagal parametrus galimybę įmanoma per išplėstinę sąsają. Pvz. Google leidžia ieškoti tik tinklapyje ("site:"), rasti nuorodas į tinklapį ("link:"), ieškoti pagal šalį, datą, domeną ir t.t. Užklausos įvedimo lauke galima įvesti ir matematines formules, naudojant Google kaip kalkuliatorių, o tai, aišku, paryškina šios sistemos savitumą: pvz. užklausai „4^2“ bus gautas rezultatas „16“.

Paieška pagal panašumą

Kiekvienas, sistemos nuomone, reikšmingas žodis reitinguojamas pagal tam tikrą kriterijų, iš reikšmingiausių žodžių automatiškai formuojama užklausa vertinama kaip naujas paieškos kriterijus.

Paieška pagal atkirtimą

Kai kurios sistemos laiko visus žodžius dešiniaisiais atkirtimais, o kai kurios išvis to nepalaiko (Google, Alltheweb). Tačiau daugumoje sistemų norint užmaskuoti dešinįjį žodžio atkirtimą užtenka panaudoti simbolį „*“.

Paieška pagal laukus (filetype:, url:, site:)

Leidžia apriboti paieškos diapazoną iki paieškos pagal URL reikšmes, pagal datas, pavadinimus ir t.t.

2.3.5. Užklausų patikslinimai

Kai kada logiškai suformuotai užklausai pateikiami tūkstančiai dokumentų, mažai susijusių su informaciniu poreikiu. Tokiu atveju gelbsti paieškos sistemos savybė nekeičiant ankstesnės užklausos logikos daryti užklausą jos rezultatui.

2.3.6. Dokumentų reitingavimas

Dokumentų reitingavimas - tai viena pagrindinių šiuolaikinių paieškos sistemų charakteristikų. Reitinguoti galima pagal sukūrimo datą, dokumentų atnaujinimą, pagal reikšmingumą (daugelis paieškos sistemų vertina dokumentų reikšmingumą pagal svorio kriterijus arba nuorodų į jas kiekį, t.y pagal citavimą). Reitingavimas pagal citavimo indeksą analogiškas mokslinių publikacijų vertinimui tradicinėje mokslinėje sferoje - PageRank (PR) pirmiausia panaudojo Google, pademonstravusia tokio metodo web-erdvei efektyvumą [23]. Po Google panašiais algoritmais ėmė naudotis kitos paieškos sistemos.

2.4. Paieškos modeliai

Šiuo metu naudojama keletas informacijos pateikimo duomenų bazėse metodų šios informacijos tolesnei paieškai. Tie metodai yra: Susietų žodžių lentelės (Word Association), Tikimybinis modelis (Probabilistic Retrieval), Matricinis latentinis-semantinis metodas (Latent Semantic Indexing), tačiau jų aprašymas – atskiro mokslinio darbo tema. Peržiūrėsime du populiariausius metodus. Pirmas paremtas aibių teorija (Boolean Retrieval), a antrasis – vektorine algebra (Vector Space Model). Abu metodai praktikoje efektyvūs, tačiau abu turi bendrą trūkumą: dokumento prasmė, pagrindinis jo turinys nustatomas pagal raktinius žodžius – jo terminus ir sąvokas, o tai iš dalies keičia turinio stilistinius atspalvius. Užtat tai leidžia atlikti greitą paiešką ir sugrupuoti dokumentus pagal formalius požymius. Šiandien šie metodai patys populiariausi.

2.4.1. LPM - Loginės paieškos modelis (Boolean Retrieval)

LPM yra klasikinis ir plačiai naudojamas informacijos pateikimo modelis, besiremiantis aibių teorija [39]. Šio modelio populiarumą nulemia jo realizacijos paprastumas, leidžiantis indeksuoti ir atlikti paiešką didelės apimties dokumentų masyvuose. Šiame modelyje dokumentai ir užklausos pateikiamos kaip raktinių žodžių (termų) morfeminių pagrindų aibė.

Tegul dokumentų masyvas C susideda iš aibės dokumentų $d_x,...,d_n$, o dokumentas d_i turi skirtingų termų aibę $T(d_i)$. Tada masyvo C žodynas T apskaičiuojamas pagal formulę:

 (3)

Šiame modelyje vartotojo užklausa yra loginis užrašas, kuriame termai susieti loginiais operatoriais AND, OR ir NOT. Naudojantis loginiu modeliu, paieškos sistemos duomenų bazė susideda iš indekso, organizuoto kaip invertuotas masyvas, kuriame kiekvienam termui iš DB žodyno egzistuoja dokumentų sąrašas, kuriame tas termas randamas.

Indekse taip pat talpinami duotojo termo radimo kiekviename dokumente dažnių reikšmės, o tai leidžia rūšiuoti sąrašą pagal radimo dažnio mažėjimą. Klasikinė DB, atitinkanti LPM, organizuota tokiu būdu, kad pagal kiekvieną termą būtų galima greitai pasiekti reikalingų dokumentų sąrašą. Be to, invertuoto masyvo struktūra leidžia greitai jį modifikuoti talpinant į DB naujus dokumentus, todėl invertuotas masyvas dažnai realizuojamas kaip B-medis.

2.4.2. VEM - Vektorinės erdvės modelis (Vector Space Model)

Šiame modelyje dokumentas nusakomas euklidinės erdvės vektoriumi, kuriame kiekvienam dokumente vartojamam termui priskiriamas jį atitinkantis svorio koeficientas. Koeficientas skaičiuojamas pagal jo radimo dokumente ar dokumentų masyve statistinę informaciją. Užklausa, kuri atitinka vartotojo tematiką, irgi yra vektorius toje pačioje euklidinėje termų erdvėje. Užklausos ir dokumento artumo skaičiavimui naudojamas atitinkamų vektorių (tematikos ir dokumento) skaliarinė sandauga [39].

VEM modelyje prie kiekvieno termo t_i dokumente d_j (ir užklausoje q) derinamas tam tikras teigiamas svoris w_{ij}. Kiekvienas dokumentas ir kiekviena užklausa gali būti išreikšti kaip k-dimensijų vektorius $||w_{ij}||_{i=1...k}$, kur k – bendras skirtingų termų skaičius visuose dokumentuose. Pagal šį modelį, dokumento d_i artumas su užklausa q vertinamas kaip jų apibūdinimo vektorių koreliacija. Koreliacija skaičiuojama kaip atitinkamų apibūdinimo vektorių skaliarinė sandauga. Atskirų termų svorio koeficientai skaičiuojami pagal formulę:

 (4)

Čia n_j — dokumentų, kuriose naudojamas termas t_j, kiekis, o N — bendras dokumentų kiekis masyve. Paprastai svorių w_{ij} reikšmes normuojamos.

Dviejų dokumentų arba dokumento ir užklausos teminiam artumui nustatyti šiame modelyje naudojama paprasta dviejų vektorių $||w_{i1}||_{i=1...k}$, ir $||w_{i2}||_{i=1...k.}$ skaliarinė sandauga $sim(d_1,d_2)\in[0,1]$. Kuo didesnis šitas dydis – tuo artimesni dokumentai.

Sistemos, sukurtos VEM pagrindu, turi tokias galimybes:

- Neriboto dydžio užklausų apdorojimas.
- Paprasta panašių dokumentų paieškos režimo realizacija.
- Paieškos rezultatų išsaugojimas virtualiame masyve tolesnei patikslintai hypernuorodų paieškai.

2.5. Paieškos sistemų darbo principai

Paieškos sistemos kūrimas susiduria su daugeliu sunkumų. Greita puslapių peržiūrėjimo technologija reikalinga puslapiams surinkti ir naujausiai informacijai saugoti. Saugojimo erdvė turi būti išnaudojama efektyviai. Indeksavimo sistema turi apdoroti šimtus gigabaitų informacijos. Užklausos turi būti vykdomos greitai, per šimtąsias ar tūkstantąsias sekundės dalis.

Paieškos sistemos susideda iš penkių atskirų komponentų [34]:

- **Spider** („voras“): programa, kuri skaito tinklapius.
- **Crawler** („keliautojas“): voras,kuris automatiškai keliauja pagal tinklapio nuorodas.
- **Indexer** (”indeksatorius“): programa, kuri analizuoja gautus iš voro tinklapius.
- **Database** (”duomenų bazė“): čia saugojami voro perskaityti tinklapiai.
- **Results engine** (”rezultatų pateikimo sistema“): paima rezultatus iš DB ir pateikia juos vartotojui.

Spider: programa, kuri skaito tinklapius. Jis dirba tokiu pat pagrindu kaip paprasta naršyklė, kai vartotojas jungiasi prie tinklapio. Voras neturi jokių vizualinių komponentų. Ta patį veiksmą (skaitymą) galima stebėti, kai su naršykle peržiūrimas tinklapio kodas („View source“). Tinklo „voro" paleidimas - sudėtinga užduotis. Yra keblių patikimumo ir produktyvaus darbo problemų ir - dar svarbiau - yra socialinių problemų. Ieškojimas ir skaitymas yra pati trapiausia programa, kadangi ji interaktyviai bendrauja su šimtais tūkstančių Interneto serverių. Voras turi būti patikimas, greitas ir robustiškas.

Crawler: Jo tikslas – nustatyti, kur toliau turi keliauti voras, remdamasis nuorodomis arba iš anksto paruoštu adresų sąrašu. Dažnai reikia uždrausti botui pasiekti konkrečias tinklapio dalis. Tam tikslui skirtas neoficialus standartas „Standart for Robot Exclusion (SRE)“, kurio remiasi beveik visi paieškos sistemų kūrėjai. Pagal šį standartą reikia tik sukurti failą

“**robots.txt**“ serverio šakniniame kataloge ir įrašyti į jį keletą apribojimo komandų. Pagal serverio log-failą galima stebėti, kokias tinklapio dalis lanko botas. Kai kurie botai identifikuojami pagal savo pavadinimą – pvz. Google "Googlebot". Kiti labiau paslėpti ir maskuojami kaip naršyklės.

Indexer: Indeksatorius dalina puslapį į skirtingas dalis (pavadinimai, nuorodos, tekstas, struktūriniai elementai, stiliaus elementai ir t.t) ir analizuoja juos. Botai nesupranta freimų, flash-animacijos ir skriptų.

DB: Tai tinklapių, kuriuos paieškos sistema indeksuoja, saugykla. Jeigu tinklapis pateko į DB, tai botai periodiškai lankys jį ir stebės visus pakeitimus. Apsilankymų kiekis priklauso nuo konkrečios paieškos sistemos.

Search Engine Results Engine: Rezultatų pateikimo sistema paima rezultatus (SERP, search engine result pages) iš DB ir pateikia juos vartotojui.

2.6. Paieškos sistema Google

“*If your page is not indexed by Google, your page does not exist on the Web“*
S. Olsen [30]

Google buvo paleistas 1998 m. Stendfordo universiteto studentų Sergejaus Brino (Sergey Brin) ir Lari Peidžo (Larry Page), tuo laiku dirbusių prie mokslinio projekto apie prasminių elementų identifikaciją hypernuorodų struktūroje.

Šiuo metu Google laikomas svarbiausia paieškos sistema pasaulyje ir aptarnauja beveik 80% planetos paieškos užklausų. Todėl bet kuriam tinklapiui, kuriam svarbus paieškos sistemų grafikas, gyvybiškai reikalinga turėti gerą Google reitingą [13]. Google serverinė turi daugiau negu 60.000 kompiuterių, o sintetinis terminas „google“ toks populiarus, kad faktiškai tapo paprastu angliškų žodžiu („to google“ = ieškoti tinkle). Bet šiandien Google - ne tik tekstinių dokumentų paieškos sistema, bet ir stambiausia vaizdų, prekių katalogų paieškos sistema, archyvuota tinklo Usenet praeitis ir dabartis, milžiniškas blog-servisas Blogger, pašto tarnyba Gmail, reklamos tarnyba AdScence ir t.t. [44]. 2005 m. kovo 31 dienos duomenimis, visame pasaulyje šioje kompanijoje visą darbo dieną dirbo 3482 darbuotojų.

2.6.1. Google sistemos architektūra:

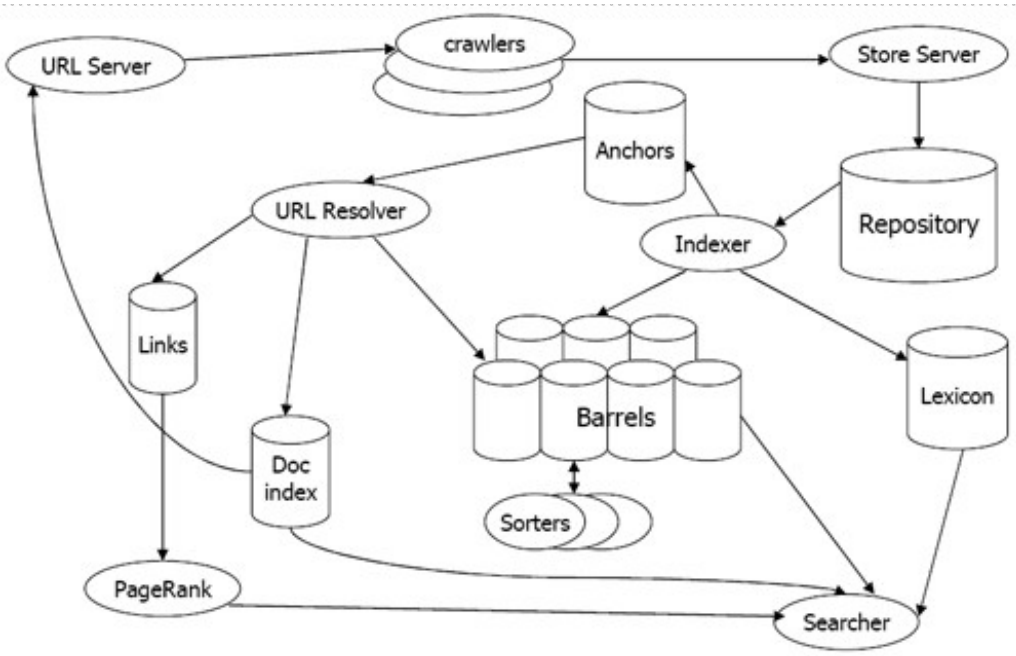
Google kodas parašytas su C++ prie Soliaris ir Linux operacinių sistemų. Google savo data-centruose naudoja pigius serverius vietoj brangių 8-procesorinių high-end serverių. Speciali failų sistema „**Google File System**“ optimizuota darbui su 64 bitų duomenų blokais. Duomenys

saugojami trimis egzemplioriais, specialūs serveriai aprūpina priėjimą prie kopijų, jeigu pagrindiniai duomenis neprieinami [20].

„Google, pvz. mes naudojame vietoj kietųjų diskų pigią DRAM atmintį – tai atrodo neįtikėtina, bet toks sprendimas pigesnis. DRAM 200.000 kartų efektyvesnė. Paieškoje daugiau vietos tobulinimui – ar tai būtų algoritmai, ar sisteminė architektūra. (Erikas Šmitas, Google vadybininkas) [38]

Siųsdamas užklausą, vartotojas patenka į geografiškai sau artimiausią Google klasterį. Išskirstyti visame pasaulyje klasteriai susideda iš kelių šimtų kompiuterių kiekvienas. Tokia architektūra leidžia išvengti perkrovų ir gina nuo katastrofiškų sistemos pertrūkių (žemės drebėjimai, globalūs elektros atjungimai) [4].

2.6.2. Google funkcionavimas ir duomenų struktūros



7 pav. ***Google funkcionavimas***

Google sistemoje puslapius skaito keletas sukurtų **„vorų" (crowlers)**. **URLserver** siunčia vorams URL sąrašus tikrinti (7 pav). Puslapiai, kurie buvo patikrinti, siunčiami į saugojimo serverį (**Store server**). Saugojimo serveris suspaudžia ir išsaugo puslapius saugykloje (**Repositary**). Kiekvienas puslapis turi jam priskirtą **docID** numerį. Indeksuotojas (**Indexer**) skaito duomenis iš saugyklos, išspaudžia dokumentus ir indeksuoja juos. Kiekvienas dokumentas išverčiamas į žodžių aibę, kuri vadinasi „pataikymas“ (**hit)**. Pataikymo įraše fiksuojama žodžių

vieta dokumente, šrifto dydis bei didžiųjų ar mažųjų raidžių parametrai. Toliau pataikymai perkeliami į „statinių“ (**barrels**) aibę ir sukuriamas dalinai rūšiuotas indeksas. Indeksuotojas atlieka ir kitą svarbią funkciją - išanalizuoja visas nuorodas kiekviename puslapyje ir išsaugo svarbią informaciją apie jas nuorodų byloje (**Anchors**). **URLresolver** skaito nuorodų bylą, paverčia santykinius URL į absoliutinius URL, grąžina juos į dokumentų indeksą (**Doc Index**), patalpina nuorodų tekstą į „statinių“ aibę. Taip pat jis naudojamas visų dokumentų reitingams suskaičiuoti. Rūšiuotojas (**Sorter**) pasiima „statines“ (**barrels**), kurios jau surūšiuotos pagal **docID**, ir perrūšiuoja pagal **wordID**, kad sugeneruotų invertuotą indeksą. **Lexicon** paima tą indekso sąrašą kartu su indeksuotojo sukurtu žodynu ir sugeneruoja naują žodyną, kuris bus naudojamas ieškotojo (**searcher**). Ieškotojas naudoja Lexicon žodyną kartu su invertuotu indeksu ir **PR**, kad atsakytų į vartotojo užklausas [8].

Google duomenų struktūros padarytos taip, kad didelės dokumentų kolekcijos galėtų būti skaitomos (crowled), indeksuojamos ir ieškomos su labai maža sąnauda. Google sukurtas vengti kreipimosi į diskus, kai tik yra įmanoma, ir tai turi didelės įtakos duomenų struktūros konstrukcijai. Pagrindinės Google duomenų struktūros yra**:**

Didelės bylos (BigFiles)

Didelės bylos - tai virtualios bylos, susiejančios daugialypes bylų sistemas ir adresuojamos 64 bitų skaičiais. Didžiosios bylos turi elementarią suspaudimo galimybę.

Saugykla (Repository)

Saugykloje saugomi pilni HTML puslapiai (8 pav). Kiekvienas puslapis suspaudžiamas naudojant **zlib** technologiją (suspaustų duomenų apimtis sumažėja beveik tris kartus).

Repositoty

sync	length	Compressed packet

Compressed packet

docID	ecode	urllen	pagelen	page

8 pav. ***Saugykla ir suspaustas paketas***

Dokumentų indeksai

Dokumentų indeksai saugo informaciją apie kiekvieną dokumentą. Informacija, saugoma kiekviename įraše, nurodo esamą dokumento būklę, nuorodas į saugyklą, dokumentų duomenis ir įvairią statistiką. Jei dokumentas jau buvo peržiūrėtas, jis taip pat turi nuorodą į kintamąją tarpinę bylą, vadinamą **docinfo**, kuriame yra URL ir dokumento antraštė. Kitu atveju nuoroda

nurodo į **URLlist,** kuriame saugomi tik URL. Šis sprendimas atsirado norint turėti pagrįstai kompaktišką duomenų struktūrą ir galimybę išrinkti įrašą vienu kreipimusi į diską paieškos metu. Be to, yra byla, kuri naudojama URL paversti į **docID**. URL sąrašas, kuris atitinka **docID**, išrikiuojamas pagal rezultatus. Kad rastume **docID** iš atskiro URL, URL rezultatinis sąrašas suskaičiuojamas ir atliekama binarinė paieška, kad būtų surastas **docID**. Šis paketinis metodas yra esminis, nes kitu atveju diskui reikėtų kreiptis į kiekvieną nuorodą, kuri reikalautų atskiro kreipimosi, ir tai truktų daugiau kaip mėnesį 322 milijonams nuorodų duomenų bazėje apdoroti.

Žodynas (Lexicon)

Žodynas susideda iš žodžių sąrašų (sujungtų kartu, bet atskirtų nuliais) ir mišrios nuorodų lentelės.

Pataikymų sąrašas (Hit list)

Pataikymų sąrašas atitinka tam tikrų žodžių tam tikrame dokumente sąrašą, įskaitant vietą, šriftą ir raidžių registro informaciją. Google kompaktiškas kodavimas naudoja du baitus kiekvienam pataikymui. Yra du pataikymų tipai: raiškusis (**fancy**) ir paprastas (**plain**). Raiškusis įtraukia savo buvimą URL, antraštėje, inkariniame tekste (**Anchor Text**) ar metatege. Paprasti įtraukia visa kita.

Paprastas pataikymas susideda iš kapitalizacijos bito, šrifto dydžio ir 12 bitų žodžio pozicijos dokumente (visos vietos, aukštesnės nei 4095, žymimos 4096). Šriftų dydis pristatomas naudojant tris bitus (tik 7 reikšmės iš tikrųjų naudojamos, nes 111 yra žymė, signalizuojanti raiškųjį pataikymą). Raiškusis pataikymas susideda iš kapitalizacijos bito, šrifto dydis nustatomas 7 bitais, 4 bitai - užkoduoti tipą ir 8 bitai - pozicijai.

Inkarinis tekstas (Anchor text)

Google susieja nuorodą ir su puslapiu, į kurį ta nuoroda nukreipta, nes nuorodos dažnai pateikia daug tikslesnius puslapių apibūdinimus negu patys puslapiai. Tai šiek tiek riboja frazių ieškojimą, kadangi nėra daug inkarų atskiriems žodžiams. Google naudoja šrifto dydžio susiejimą su visu dokumentu, nes kai vykdoma paieška, nenorima išrinkti, kitaip identiškas dokumentas atrodys skirtingai tik todėl, kad jo šrifto dydis didesnis ar mažesnis.

Pirminis indeksas (Forward Index)

Pirminis indeksas jau dalinai suindeksuotas. Jis saugomas „statinėse“ (**barrels**). Kiekviena statinė saugo wordID sritį. Jei dokumentas savyje turi žodžių, įeinančių į tam tikrą paketą, docID įrašomas į paketą atsižvelgiant į wordID sąrašą su pataikymų sąrašais, kurie atitinka tuos

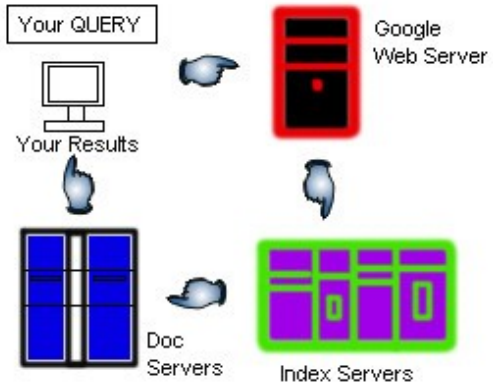
žodžius. Ši schema reikalauja šiek tiek daugiau atminties dėl docID atkartojimo; skirtumas labai nežymus ir sutaupo daug laiko galutinėje indeksavimo fazėje, kurią atlieka rūšiuotojas. Dar daugiau: užuot saugojusi tikrąjį wordID, Google saugo kiekvieną wordID, kaip susijusį su wordID minimumu, kuris įeina į paketą, kuriame wordID saugomas. Tokiu būdu galima panaudoti tik 24 bitus wordID ir nerūšiuotiems paketams, paliekant 8 bitus pataikymų sąrašų ilgiui.

Invertuotas indeksas (Inverted Index)

Invertuotas indeksas susideda iš tų pačių paketų, kaip ir pirminiame indekse, išskyrus tai, kad jie buvo perdirbti iš rūšiuotojo. Kiekvienam galiojančiam wordID žodynas saugo nuorodą, kuri nurodo paketą, kuriame yra wordID. Jei nurodoma į dokumentų sąrašą iš docID - kartu su jais nurodomas ir atitinkamas pataikymų sąrašas. Šis dokumentų sąrašas pristato visas to žodžio buvimo vietas visuose dokumentuose. Svarbi problema yra tai, kokia tvarka docID turėtų pasirodyti dokumentų sąraše. Viena paprasta išeitis - saugoti juos išrūšiuotus pagal docID. Tai leistų greitą skirtingų dokumentų sąrašų suliejimą sudėtinėse žodžių užklausose. Kita galimybė - saugoti juos išrūšiuotus pagal žodžio pasirodymo dokumente reitingą. Tai leidžia atsakyti į vieno žodžio užklausas ir sukuria tikimybę, kad atsakymas į sudėtines žodžių užklausas būtų pradžioje. Kaip bebūtų, suliejimas daug sudėtingesnis. Taip pat tai padaro daug sudėtingesnį ir vystymą, nes pokyčiai reitingavime reikalauja perkurti indeksą.

2.6.3. Užklausos apdorojimas

Užklausos apdorojimo procesą galima pavaizduoti schema (9 pav.) [8]:



9 pav. ***Užklausos apdorojimo procesas***

Index Serveriai turi invertuotą indeksą, kuris suderina kiekvieną užklausos žodį su atitinkamų dokumentų sąrašu (hit list). Toliau, sukryžiuojant žodžių sutapimų sąrašus, nustatoma relevantiškų dokumentų identifikatorių eilė, ir kiekvienam dokumentui suskaičiuojamas relevantiškumo laipsnis (svoris). Svoris nustato rezultatų išvedimo tvarką. Paieškos procesas yra labai didelės apimties: nesuspausti dokumentai ir indeksai užima keletą dešimčių terabaitų. Todėl paieška vykdoma lygiagrečiai, dalijant indeksą į „nuolaužas“ (shards), kurių kiekviena apdoroja atsitiktinai pasirinktą indekso dokumentų poaibį. Kiekvieną „nuolaužą“ aptarnauja atskiras klasteris. Jeigu klasteris neprieinamas, balansatorius perkels skaičiavimus į kitą klasterį. Užklausos atlikimo galutinės fazės įvykdymo rezultatas – tai – sutvarkytas dokumentų identifikatorių (docIDs) sąrašas. [4]

Doc Serveriai pagal dokumentų identifikatorių (docIDs) sąrašą nustato URL, tinklapio aprašymą ir visą reikalingą rezultatų sąrašo išvedimui informaciją. Šių skaičiavimų spartinimui kreipiamasi ne į patį tinklą, o į globaliojo tinklo kopiją, paskirstytą pagal klasterius.

Papildomai prie šių fazių Google taip pat atlieka keletą pagalbinių uždavinių: sintaksės tikrinimą, relevantiškos reklamos kūrimą. Kada visos fazės įvykdytos, Google generuoja atitinkamą HTML kodą ir grąžina jį vartotojui.

2.6.4. Robotas (botas) Googlebot

Šimtams milijonų Interneto puslapių reitinguoti Google turi greitai dirbančią tyrimo sistemą. Kiekvienas URLserveris pateikia URL sąrašus tam tikram „vorų" skaičiui (Google tipiškai naudojo 3-4 vorus). Abu: ir URLserveris, ir „vorai“, yra įdiegti į Python. Kiekvienas „voras" laiko apytiksliai 300 atvirų jungčių vienu metu. Tai būtina, norint gauti puslapius pakankamai greitai. Ribiniais greičiais sistema gali patikrinti iki 100 tinklapių per sekundę, naudodama keturis „vorus". Tai sudaro apytiksliai 600Kb duomenų per sekundę. Pagrindinis efektyvumo kriterijus tenka DNS paieškai. Kiekvienas „voras" prižiūri savo nuosavus DNS atsargos resursus, taigi nebereikia vykdyti DNS paieškos, prieš tikrinant kiekvieną dokumentą.

Kiekviena iš šimtų jungčių gali būti skirtingose būsenose: ieškoti DNS, jungtis prie tinklo, siųsti užklausas, gauti atsakymus. Voras naudoja asinchroninius Į/I (I/O) įvykiams valdyti, ir daug užklausų tinklapiui keliama iš vienos būsenos į kitą būseną lygiagrečiai.

Google apeina (crawls) tinklapius dviem būdais:

- **Main Crawl** (crawl1) iš IP 216.239.46.xxx
- **Fresh Crawl** (crawl2) iš IP 64.68.82.xxx

Paprastai naują tinklapį pirmiausia aplanko voras **crawl2**, nors iš karto po atnaujinimo tinklapis iš pradžių gali būti aplankytas voro **crawl1**.

Naujas tinklapis nepateks į Google pagrindinį indeksą iki tol, kol:

- jo neaplankys voras **Crawl1**
- jis „išgyvens“ atnaujinimą po **Crawl1** apsilankymo.

Kol šitai neįvyks, bet koks tinklapio gautas reitingas greičiausiai bus laikinas. Apytiksliai naujo tinklapio gyvavimo ciklas yra toks:

Fresh Crawl → atnaujinimas → **Main Crawl → atnaujinimas** → pagrindinis indeksas

- Per mėnesį tarp atnaujinimų vykdomas Fresh Crawl ir voras **crawl2** aplanko naują tinklapį. Naujas tinklapis vertinamas ir įterpiamas į paieškos rezultatus, net jeigu jis nebuvo įterptas į atnaujinimą ir neįdėtas į pagrindinį Google indeksą. Tinklapio pozicijos nestabilios ir labai dažnai keičiasi.
- Paskui ateina atnaujinimo laikas. Naujas tinklapis dar nepatenka į pagrindinį indeksą. Kada atnaujinimas baigiamas, tinklapis vis dar laikomas nauju tinklapiu ir Google dar nerodo visų nuorodų į jį.
- Greitai po to vykdomas Main Crawl ir tinklapį aplanko voras **crawl1**. Dabar tinklapiui reikia sulaukti kito atnaujinimo, kad patektų į pagrindinį indeksą. Iki tol tinklapis vis dar laikomas nauju tinklapiu.
- Po atnaujinimo puslapis patenka į pagrindinį indeksą.
- Google indeksuoja tik pirmus dokumento 101Kb.

Privilioti robotą į tinklapį gali:

- Didelis kiekis nuorodų iš kitų tinklapių.
- Aukštas tinklapio PageRank reitingas.
- Rankinis URL įterpimas per specialią formą (http://www.google.com/addurl.html).
- Tinklapio peržiūrėjimas naršyklėje, kur įdėta GoogleToolbar įrankių juosta (tai neoficiali informacija, tačiau autoriaus tyrimai tą faktą patvirtino).

Kaip robotas mato tinklapį: http://www.seo-browser.com/

Kitų paieškos sistemų botų sąrašas pateiktas 2-ame Priede.

2.6.5. Google Šokis („Google Dance“)

Google šokis – tai laiko periodas, kada Google perskaičiuoja savo reitingus, o paieškos rezultatai smarkiai kinta 3-5 dienas. Šitas globalus atnaujinimas vyksta vidutiniškai kas 36 dienos (10 kartų per metus).

Kadangi Google, kaip ir bet kuri kita paieškos sistema, priklauso nuo vartotojų ir turi dirbti 24 valandas per parą, septynias dienas per savaitę, tai atnaujinimas tampa rimta problema. Google serveriai (2 lentelė) negali būti sustabdyti techniniam aptarnavimui ir negali būti išjungti iš tinklo net ir minutei.

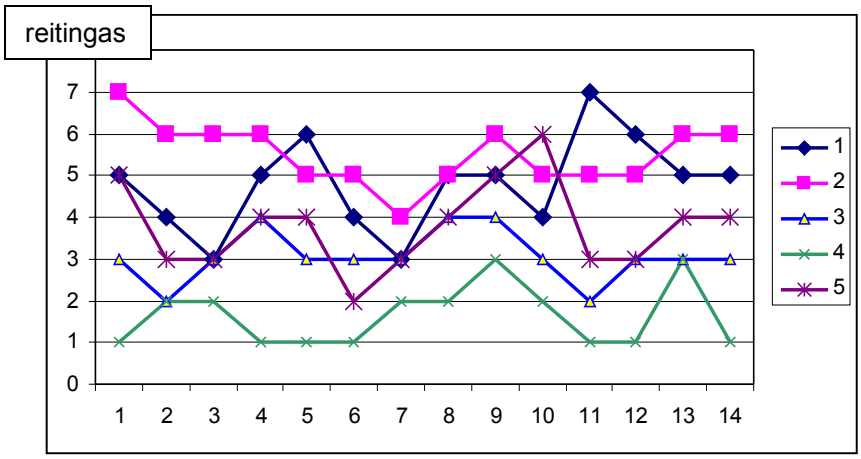
2 lentelė. ***13 pagrindinių Google serverių pavadinimai ir IP adresai***

www-ex.google.com	216.239.33.100
www-sj.google.com	216.239.35.100
www-va.google.com	216.239.37.100
www-dc.google.com	216.239.39.100
www-ab.google.com	216.239.51.100
www-in.google.com	216.239.53.100
www-zu.google.com	216.239.55.100
www-cw.google.com	216.239.57.100
www-fi.google.com	216.239.41.100
www-gv.google.com	216.239.59.100
www-kr.google.com	66.102.11.100
www-mc.google.com	66.102.7.100
www-lm.google.com	66.102.9.100

Reitingo perskaičiavimo metu įdomu patikrinti skirtingus Google serverius: jie parodys skirtingus, dažnai net labai skirtingus rezultatus, ir galima sakyti, kad jie tarsi „šoka“ – iš čia ir atnaujinimo pavadinimas - **Google Šokis** [36].

2.6.6. Autoriaus stebėjimai: Kasdieninis reitingų perskaičiavimas

Tas, kuris nuolat naudojasi Google, tikriausiai pastebėjo, kad net vienas ir tas pats tinklapis skirtingu laiku gali būti skirtingose pozicijose. Šitai Google savybei paaiškinti buvo atlikti tyrimai, kurių pagrindu galima tvirtinti, kad Google rečiausiai kasdieną perskaičiuoja tinklapių reitingus (pavadinsime tai „mini-shift“; kadangi toks perskaičiavimas skiriasi nuo Google šokio, kada globaliai perskaičiuojami tinklapių PR). Stebėjimas vyko naudojant 5 autoriaus sukurtus tinklapius 14 dienų (15/04/2005-29/04/2005) ir kaip rezultatas buvo gautas grafikas (10 pav.) ir lentelė (3 lentelė):



10 pav. ***Google kasdieną perskaičiuoja tinklapių reitingus***

3 lentelė. ***Eksperimento tinklapiai ir užklausos***

№	Tinklapis	Užklausa (raktinis žodis)
1	www.hbh.lt	alaus darykla
2	www.mamarosa.lt	viešbutis palanga
3	www.termionas.lt	butas klaipeda
4	www.notarius.lt	notaras klaipeda
5	www.siampeksas.lt	galvos apsauga

Reitingų poslinkius galima paaiškinti naujų tinklapių su duotais raktiniais žodžiais pasirodymu ir jau egzistuojančių tinklapių pasikeitimais ir atnaujinimais.

2.7. Paieškos sistemų algoritmų ir jų pakeitimų stebėjimas.

Be jau minėtų 2.6.6 skyrelyje stebėjimų autorius atliko sistemingą Google paieškos monitoringą ir patvirtino tokius faktus:

Google taiso klaidas paieškos rezultatuose.

Tokią Google savybę autorius pastebėjo, kai vykdė paiešką pagal raktinius žodžius „Search Engine Optimisation“. Gautuose rezultatuose buvo ir žodžiai „optimi**s**ation“ su „s“ – ir žodžiai „optimi**z**ation“ su „z“. Paieška pagal žodžius „e-mail solution“ taip pat davė rezultatus su „email solution“, o „e-commerce“ – „ecommerce“.

Google analizuoja nuorodų natūralumą

Nuo 2005 m. balandžio mėn. Google paieškos rezultatai smarkiai pasikeitė. Dabar, kada dokumente pasirodo naujos nuorodos, Google tikrina, kaip jų pasirodymas arba išnykimas veikia kitas nuorodas, susijusias su dokumentu, ir dokumento releventiškumą. Kartu analizuojama

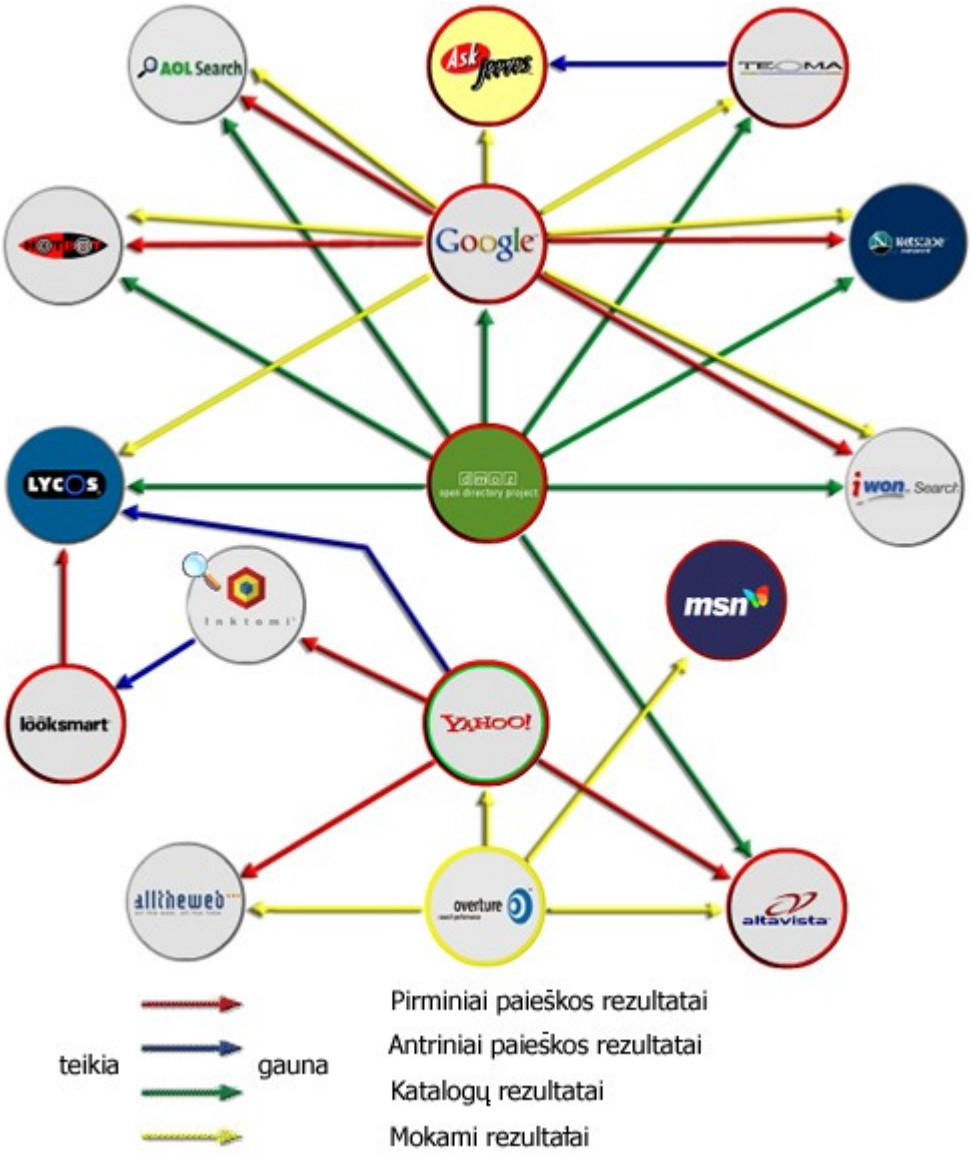
daugybė skirtingų parametrų, kurie leidžia Google apskaičiuoti nuorodos kokybę ir jos pasirodymo tinklapyje natūralumą. Dabar Google seka nuorodos pasirodymo datą ir analizuoja nuorodas, atsižvelgdamas į jų pasirodymo periodiškumą, ir ieško tendencijų.

Google naudoja morfologinę raktinių žodžių analizę paieškos rezultatuose.

Šiuo metu Google siūlo keletą relevantiškų raktinių žodžių variantų atliekant paiešką. Pvz.: sodas, sodai, sodininkystė. Optimizacijos atžvilgiu, tai leidžia raktinių žodžių srityje konkuruoti su didesniu tinklapių kiekiu [24].

2.8. Ryšys tarp paieškos sistemų

Be Google, pasaulyje egzistuoja daug paieškos sistemų, tarp kurių yra apie 16 stambiausių (Yahoo!, AltaVista, AOL, Excite, FindWhat, LookSmart, Lycos, MSN, DMOZ, Overture, iWon ir t.t). Visos jos susijusios tarpusavyje apsikeitimų paieškos rezultatais pagrindu Kai kurios turi bendrą savininką ir suprojektuotos dirbti palaikydamos viena kitą. Detaliau sistemų ryšį galima panagrinėti 11 pav. Jame pateiktos tų 16 stambiausių sistemų bei skirtingo pobūdžio ryšiai tarp jų [9]. **Pirminiai rezultatai** ateina iš paieškos sistemos savos duomenų bazės. **Antriniai rezultatai**, kuriuos pateikia sistemos, – tai rezultatai iš sistemų-partnerių.



11 pav. ***Ryšys tarp paieškos sistemų.***

2.9. Metapaieškos sistemos

Kiekviena paieškos sistema turi tik savo resursų ribotą, pasiekiamų paieškai dokumentų aibę. Nė viena iš tokių paieškos sistemų negali apimti visų resursų. Problemai išspręsti ir paieškos galimybėms išplėsti sukurtos sistemos, vadinamos metapaieškos sistemomis.

Metapaieškos sistemos (Metasearch Engines, MSE) – tai paieškos sistemų antstatai, kurie neturi savo DB ir, gavę iš vartotojo užklausą, savo ruožtu formuoja naują užklausą kelioms išorinėms paieškos sistemoms, o po to pagal savo algoritmą analizuoja gautus rezultatus ir pateikia nuorodų sąrašą, surūšiuotą pagal kelių sistemų pateiktus reitingus [40]. Stambiausios metapaieškos sistemos yra:

- **MetaCrawler**: www.metacrawler.com

  Naudoja rezultatus iš: AltaVista, Direct Hit, Excite, FindWhat, Kanoodle.com, LookSmart, DMOZ, Overture, Sprinks, WebCrawler)
- **Mamma**: www.mamma.com

  Naudoja rezultatus iš: 7Search.com, About.com, Ah-ha, BIGwhat.com, Direct Hit (Ask Jeeves), Kanoodle.com, LookSmart, Lycos, Mamma Collections, MSN, Overture.com, Sprinks
- **Ixquick**: www.ixquick.com

  Naudoja rezultatus iš: AllTheWeb, AltaVista, AOL, Excite, FindWhat, LookSmart, Lycos, MSN, DMOZ, Overture, Sprinks, Yahoo
- **DogPile!:** www.dogpile.com

  Naudoja rezultatus iš: Ah-ha, AltaVista, Direct Hit, Dogpile Web Catalog, ePilot.com, FindWhat, Kanoodle, LookSmart, DMOZ, Overture, RealNames, SearchHippo.com, Sprinks, Yahoo!

Tačiau labai dažnai savas relevantiškumo nustatymo algoritmas pateikia blogesnius rezultatus, negu paieškos sistemos, į kurias metasistema kreipiasi. Todėl laiko sąnaudų ekonomija yra nežymi [34].

2.10. Specializuotos paieškos sistemos

Specializuotos ar teminės paieškos sistemos praktiškai apima visus įmanomus dalykus. Yra daug specializuotų paieškos sistemų, skirtų astronomijai, politikai, žvejybai ir t.t. Dėl siauros panašių sistemų specializacijos paieškos rezultatams būdingas labai aukštas relevantiškumas, todėl tokių sistemų egzistavimas labai naudingas vartotojams.

Specializuotų paieškos sistemų paieškos sistema randama šiuo adresu:
http://www.searchengineguide.com/searchengines.html

2.11. Internetinių informacinių resursų katalogai (directories)

Terminas "paieškos sistema" dažnai vartojamas bendrai apibūdinti tiek „crowler“ tipo sistemoms, tiek žmogaus generuojamiems katalogams. Šie du paieškos sistemų tipai surenka ir pateikia duomenis radikaliai skirtingais būdais.

Internetiniai katalogai pasirodė anksčiau, negu tikrosios paieškos sistemos, kadangi jie psichologiškai artimesni žmogaus mąstymui ir tradiciniams katalogams bibliotekose. Jie susideda iš teminių kategorijų, rūšiuojami pagal abėcėlę ir nesinaudoja relevantiškumo principu. Nustatant tinklapio kokybę, didelį vaidmenį turi subjektyvūs įspūdžiai redaktoriaus, atsakingo už tinklapio įtraukimą į katalogą. Žinomiausi katalogai yra:

Pasauliniai:

Yahoo!: http://www.yahoo.com

DMOZ: http://www.dmoz.org

LookSmart: http://www.looksmart.com/

NBCi: http://nbci.msnbc.com/

Lietuviški:

Takas A-Z: http://www.takas.lt/a-z/a-z.php

Delfi Web A-Z: http://www.delfi.lt/directory/

NetSprint: http://www.seklys.lt/

Search.Lt: http://www.search.lt/

Tinklapių Katalogas: http://www.katalogas.net/

Katalogas **DMOZ ODP** (The Open Directory Project) šiandien yra pats didžiausias pasaulyje. Praktiškai visos paieškos sistemos paima duomenis iš šito katalogo. Gerai žinomas katalogas „Google Directory“ praktiškai yra DMOZ kopija, tik su tuo skirtumu, kad nuorodos jame yra rūšiuojamos pagal relevantiškumą, o ne pagal abėcėlę. Google Directory atnaujinama rečiau, negu DMOZ, todėl jeigu tinklapis pateko į DMOZ, tai Google Directory kataloge jis atsiras per 3-4 mėnesius. Registracija DMOZ‘e suteikia esmingą PR priaugimą, todėl tinklapio įtraukimas į DMOZ labai reikšmingas.

2.12. Tinklapių registracija kataloguose ir paieškos sistemose

Registracija vyksta pačiame paieškos serveryje. Žemiau pateiktos nuorodos į pačių populiariausių paieškos sistemų registracijos puslapius (4 lentelė):

4 lentelė. ***Populiariausių paieškos sistemų registracijos puslapių adresai***

AOL:	http://search.aol.com/add.adp
Alta Vista:	http://www.altavista.com/sites/search/addurl
Excite:	http://www.excite.com/info/add_url
All The Web:	http://www.alltheweb.com/add_url.php
Overture:	http://www.overture.com/d/about/advertisers/
Google:	http://www.google.com/addurl.html
HotBot:	http://hotbot.lycos.com/addurl.asp
Inktomi:	http://www.positiontech.com
LookSmart:	http://submit.looksmart.com/info.jhtml
Lycos:	http://insite.lycos.com/searchservices/
Open Directory:	http://dmoz.org/add.html
Northern Light:	http://www.northernlight.com/docs/regurl_help.html
Yahoo!:	http://docs.yahoo.com/info/suggest/

Anksčiau ir net dabar visur galima pamatyti skelbimus, siūlančius užregistruoti tinklapį daugelyje paieškos sistemų. Bet paieškos sistemos pradėjo laikyti automatinę registracija spamu ir uždraudė ją arba ignoruoja panašiu būdu registruotus tinklapius. Kai kurios paieškos sistemos pareikalavo įvesti slaptažodį, vaizduojamą kaip paveikslėlį registracijos puslapyje, tokiu būdu apsigindamos nuo robotų [3]. Todėl rankinė tinklapių registracija tapo įprasta praktika.

Mitas apie 50.000 paieškos sistemų

Egzistuoja daugybė tarnybų, kurios žada užregistruoti tinklapį 50.000 paieškos sistemų. Kai kurios tinklapių prastūmimo programos irgi teigia, kad užregistruos tinklapį 1000 paieškos sistemų. Bet Internete tiesiog nėra tūkstančio paieškos sistemų! Beveik visos iš tų 50.000 „sistemų“ yra paprastai FFA (Free-Far-All) tinklapiai – t.y nuorodų fermos, kuriose nuolat vyksta nuorodų rotacija, ir paieškos sistemos laiko tokius tinklapius spamu. Registruodamas tokiose sistemose savo tinklapį, vartotojas gauna tūkstančius laiškų su spamu ir nieko daugiau. Be to, kažin ar tinklapis apie elektrinius variklius sulauks bent vieno lankytojo iš Renesanso epochos menininkų katalogo...

Mokama registracija (PPC, pay-per-click)

Mokama registracija yra nemokamos registracijos alternatyva. Yra daug paieškos sistemų, kurios ima mokestį už vartotojo tinklapio nuorodos patalpinimą į savo indeksą. Tokių sistemų pliusas yra tas, kad jie tikrai registruoja tinklapį ir daro tai labai greitai. Minusas yra tas faktas, kad pirmieji reitinge pasirodo tie tinklapiai, už kuriuos užmokėta daugiausia.

3. TINKLAPIŲ REITINGAVIMO ALGORITMAI

3.1. Citavimo indeksas

Informacijos apimties augimas Internete sukūrė paieškos sistemoms problemą: relevantiškų užklausai dokumentų kiekis nuolat auga ir korektiškai reitinguoti tinklapius tapo labai sudėtinga. Be to, praeities reitingavimo algoritmai buvo bejėgiai prieš paprastą apgaulę: nukopijavus raktinių žodžių struktūrą iš tinklapio su aukštu reitingu buvo galima sulaukti puikaus rezultato. Atsirado būtinybė atskirti „autoritetingą“ informaciją nuo „neautoritetingos“. Paaiškėjo, kad padėti nustatyti autoritetingumą gali dokumentų nuorodų topologija – t.y dokumento citavimo indekso apskaičiavimas (kiek nuorodų iš kitų tinklapių veda į šį tinklapį).

3.2. Puslapio reitingas PageRank

The ranking function has many parameters like the type-weights and the type-prox-weights
Figuring out the right values for these parameters is something of a black art.
Brin S., Page L. “The Anatomy of a large scale hypertextual web search engine”

PageRank tipo algoritmai, naudojantys citavimo indekso skaičiavimo algoritmus, paplito po to, kai 1996 m. Stendfordo universiteto studentai Sergejus Brinas (Sergey Brin) ir Lari Peidžas (Larry Page) pirmieji pritaikė šio tipo algoritmą savo moksliniame projekte “Goolge”. Google PR algoritmas, skirtingai nuo paprasto citavimo indekso, nelaiko visų nuorodų lygiomis ir skaičiuoja jų svorius [2]. Google teigia, kad PageRank algoritmas susideda iš 500 milijonų kintamųjų ir 2 milijardų sąlygų [8].

Puslapių reitingas gali būti laikomas atsitiktinio vartotojo („random surfer“) elgsenos modeliu. Tarsime, kad yra „atsitiktinis vartotojas“, kuris atsitiktinai naršo po atsitiktinį puslapį, tiesiog spaudo nuorodas ir niekad nepaspaudžia „atgal“, tačiau jam atsibosta, ir jis pradeda naršyti po kitą atsitiktinį puslapį. Tokios elgsenos modelis – pirmos eilės Markovo grandinė („first-order Markov Chain“). Joje tikimybė, kad „atsitiktinis klaidžiotojas" aplankys puslapį, ir yra Puslapio Reitingas (PR). Iškrovimo faktorius (dumping factor) *d* yra tikimybė, kad kiekvienas puslapis nusibos „atsitiktiniam vartotojui“ ir jis paprašys kito atsitiktinio puslapio, paspaudęs nuorodą, arba grįš į puslapį, iš kurio atėjo, su tikimybe *(1-d)* “normalizacijos konstanta“ [15]. Konstanta reikalinga, jei puslapyje nėra išeinančių nuorodų (outlinks) [14]

Puslapis gali turėti aukštą Puslapio Reitingą, jei yra daug puslapių, kurie veda į jį, arba jei yra keletas puslapių, kurie nurodo į jį ir turi aukštą Puslapio Reitingą. Akivaizdu, kad puslapiai, kurie turi daug nuorodų iš daugelio vietų visame Internete, yra verti peržiūrėti. Taip pat

puslapiai, kurie tikriausiai turi vieną nuorodą iš globalių portalų (pvz. Yahoo!), taip pat verti peržiūrėti. Jei puslapis būtų nelabai kokybiškas, nelabai tikėtina, kad Yahoo svetainė turėtų nuorodą į šį puslapį. PR prižiūri abu šiuos atvejus ir viską, kas susiję, rekursiškai matuodamas svorius tarp Interneto nuorodų struktūros.

Tarkime, puslapis *u* susietas su keliais puslapiais *v*, kurie nurodo į jį. Tada puslapio *u* reitingas *pr(u)* užrašomas tokia formule [18]:

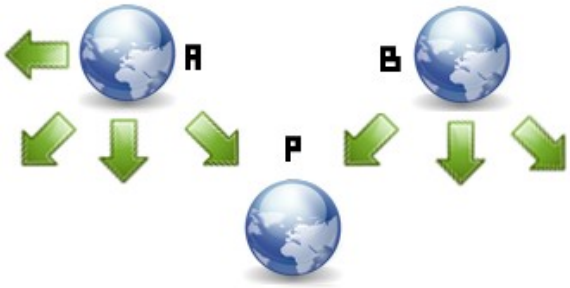
$$pr(u) = (1-d) + d \sum_{(v,u)\in E} pr(v)/outdegree(v) \tag{5}$$

Normuotas iškrovimo faktorius *d* apibrėžtas intervale [0, 1]. Google sistemoje *d* reikšmė lygi 0,85, o *n* - grandinės mazgų skaičius. „*outdegree(v)*“- išeinančių puslapio nuorodų skaičius.

Iš formulės seka:

- Kuo daugiau nuorodų iš kitų tinklapių turi tinklapis, tuo aukštesnis jo PR.
- Kuo aukštesnis tų tinklapių PR, tuo aukštesnis jo PR.
- Kuo mažiau nuorodų yra tinklapyje, to didesnį svorį jos neša.
- Kiekvienas unikalus tinklapio puslapis turi savo PR (dažnai žmonės klysta, galvodami, kad šakninio puslapio PR yra viso tinklapio PR) [41].

Pavyzdžiui: 2 populiarūs tinklapiai nurodo į trečią (12 pav.).



12 pav. ***Mazgo „P” PR bus:***

Jo PR bus skaičiuojamas pagal formulę:

$$pr(u) = (1-d) + d\left(\frac{1}{4} pr(A) + \frac{1}{3} pr(B)\right) \tag{6}$$

Iš to pavyzdžio seka, kad:

- Kuo puslapis yra arčiau šakninio puslapio, tuo daugiau svorio jis turės. O geriausiai turėti nuorodą iš šakninio (pagrindinio) puslapio, nes dažniausiai jis turi aukščiausią PR.
- Kuo geriau organizuota tinklapio vidinių nuorodų struktūra, tuo didesnį svorį turės kiekvienas tinklapio puslapis [12].

Apskaičiuoti tam tikro tinklapio PR galima pasinaudojant elektronine skaičiuokle: http://www.webworkshop.net/pagerank_calculator.php

3.3. Nuorodos ir nuorodų populiarumas

Tinklapio nuorodų populiarumas ("link popularity") – tai kitų tinklapių, kurie turi nuorodą į tinklapį, kiekis. Tai labai reikšmingas ir patikimas faktorius skaičiuojant tinklapio reitingą. Nuorodų populiarumą labai sunku sufalsifikuoti, nes tam tikslui pasiekti reikia valdyti daug skirtingų tinklapių.

3.3.1. „Smėlio dėžės" efektas

Asmeniniai tyrimai parodė, kad vienas ir gana svarbus reitingo skaičiavimo parametras tiesiogiai priklauso nuo tinklapio egzistavimo trukmės (laikas nuo pirmojo boto apsilankymo iki dabartinio momento). Naujas, ką tik paleistas tinklapis turės labai artimą nuliui PR. Šitie stebėjimai patvirtina informaciją, gautą interviu su vienu iš Google kompanijos darbuotoju – **GoogleGuy** metu. Jis teigia, kad „naujiems tinklapiams reikia laiko norint pasiekti aukštą PR ir geras pozicijas Google reitinge" [20]. Pagal šią informaciją ir paskutinį Google patentą [1] galima tarti, kad toks „karantino" prieš stojant į reitingą algoritmas – „Smėlio dėžės" efektas („**SandBox**" **effect**) **-** iš tikrųjų egzistuoja. Karantino trukmė tiksliai nenustatyta, bet stebėjimai parodė, kad karantinas tęsiasi apie 3-6 mėnesius, o prasideda po 15-30 dienų po tinklapio pataikymo į pagrindinį indeksą. Manoma, kad karantinas reikalingas spamo tinklapiams filtruoti.

2001 m. pabaigoje Google įvedė naują baudos už įtartiną optimizacijos techniką rūšį: **PR0**. Oficialios Google informacijos apie tai nėra, sakoma tik tai, kad **PR0** bauda (už spamą ir paieškos sistemos apgaulę) – tai lengvesnė alternatyva galutiniam pašalinimui iš reitingo [37].

3.3.2. Nuorodų tipai

Egzistuoja 3 tipų nuorodos:

- **Išorinės nuorodos** (outbound links) - tinklapio nuorodos į kitus tinklapius.
- **Atbulinės nuorodos** (inbound links) - kitų tinklapių nuorodos į tinklapį.
- **Vidinės nuorodos** (insite links) - tinklapio nuorodos į kitus tinklapio vidinius puslapius.

Yra klaidinga nuomonė, kad jeigu tinklapis turi išorinių nuorodų, tai mažina jo PR. Iš tikrųjų taip nėra. Mažėja tik nuorodų svoriai, t.y. – kuo daugiau nuorodų išeina iš vieno puslapio – tuo mažiau svorio neša kiekviena iš jų. Išeinančio PR formulė yra:

(7)

Čia *n* – išeinančių nuorodų kiekis.

Autoriaus skaičiavimai ir eksperimentai su 60 tinklapių leido sudaryti apytikslių PR reikšmių priklausomybės nuo atbulinių nuorodų lentelę (5 lentelė):

5 lentelė. ***Skirtingo PR nuorodų kiekis, reikalingas pasiekti tam tikrą PR.***

Nuorodos PR	Nuorodų kiekis norint pasiekti:					
	PR3	**PR4**	**PR5**	**PR6**	**PR7**	**PR8**
1	555	3,055	16,803	92,414	508,277	2,795,522
2	101	555	3,055	16,803	92,414	508,277
3	18.5	101	555	3,055	16,803	92,414
4	3.5	18.5	101	555	3,055	16,803
5	1	3.5	18.5	101	555	3,055
6	0.5	1	3.5	18.5	101	555
7	0.5	0.5	1	3.5	18.5	101
8	0.5	0.5	0.5	1	3.5	18.5
9	0.5	0.5	0.5	0.5	1	3.5
10	0.5	0.5	0.5	0.5	0.5	1

3.3.3. Paieškos kokybė

Daugelis dažnai pastebi, kad tinklapis su žemesniu PR pagal užklausą reitinge stovi aukščiau, negu jo konkurentai su aukštesniu PR. Tai sukelia nuostabą, kadangi manoma, kad tinklapio reitingas tiesiogiai priklauso nuo PR. Bet to paaiškinimas egzistuoja [5]:

Natūralus PR – tai tik tinklapio populiarumas ir tik vienas faktorius tinklapio reitingo algoritmo skaičiavimuose. Tinklapio PR galima sužinoti su GoogleToolbar įrankių juostos

pagalba. Iš tikrųjų PR, rodomas toje juostoje (PR0-PR10), – tai tik realaus PR netiesinė aproksimacija.

13 pav. ***Google įrankių juosta (GoogleToolbar)***

Lentelėje 6 pateikti skaičiai neteisingi. Kokie tie skaičiai iš tiesų – komercinė Google paslaptis.

6 lentelė. ***Tikrojo PR ir GoogleToolbar PR apytikslio atitikimo lentelė***

Tikras PR (skaičiai neteisingi)	GoogleToolbar PR
0,00000001 - 5	1
6 - 25	2
26 - 125	3
126 - 625	4
626 - 3125	5
3126 - 15625	6
15626 - 78125	7
78126 - 390625	8
390626 - 1953125	9
1953126 +	10

Kitas labai svarbus faktorius – **svoris pagal užklausą.** Jeigu PR būtų lygus paieškos sistemos reitingų rezultatams, tai **Yahoo!** (tinklapis su pačiu aukščiausiu PR10) būtų №1 kiekvienam paieškos rezultatui. Bet taip nėra. Google nagrinėja nuorodų kontekstą ir svorio prideda tik tos nuorodos, kurios atitinka užklausą (14 pav.).

14 pav. ***Rezultatai pagal raktinį žodį „Dizainas“ (2005-05-18). PR3 turintis tinklapis aplenkė studijos „Lietaus dizainas“ tinklapį, turintį RP6***

Nepaisant to, kad PR - tai dominuojantis faktorius skaičiuojant reitingą, yra ir kitų svarbių faktorių:

- Tinklapio amžius.
- Domeno vardas.
- Failų ir katalogų pavadinimai.
- Tegas <Title>.
- META tegai <Keywords>, <Description>.
- Failas „robots.txt“ (arba jo nebuvimas).
- HTML kodo, esančio prieš tekstą, dydis.
- Klaidos kode.
- Klaidos tekste.
- Teksto apimtis.
- Raktinių žodžių dokumente kiekis.
- Raktinių žodžių atstumas nuo teksto pradžios.
- Raktinių žodžių sankaupa.
- Tikslus atitikimas raktinei frazei.
- Raktinių žodžių išskyrimas (tegai <B>, < H1 > - < H6 >, <STRONG>).
- Tego <IMG> atributas ALT.
- Tegas <A>.
- Tegas <FRAME>.
- Tegas <SCRIPT>.
- Vidinių nuorodų organizacija (Gylis, aprėpimas, kiekis).
- Tinklapių, iš kurių ateina nuorodos į tinklapį, reitingas ir atitikimas temai.
- Virtualaus serverio, kur yra patalpintas tinklapis, kokybė (grafiko apribojimai, patikimumas).
- Tinklapio atitikimas pagal temą pateiktiems jame raktiniams žodžiams
- Tinklapio informacijos atnaujinimo periodiškumas ir paskutinio atnaujinimo data.

Labai svarbūs faktoriai, kurių optimizatorius negali valdyti, yra:

- Pasirinktos tinklapio temos konkurencijos lygis.
- Tinklapio geografinė padėtis.

- Kitų pasirinktos temos resursų kokybė.
- Pasirinktos temos paklausa.
- Spamerių, atakuojančių paieškos sistemą pagal pasirinktą temą, kiekis.

Atbulines nuorodas galima peržiūrėti, parašus Google pvz.: “link:www.jusu_tinklapis.lt” [5]. Google pateiks tinklapių, kurie turi nuorodas į tinklapį, sąrašą. Į sąrašą įjungiamos tik tos nuorodos, kurios turi tam tikrą PR.

Nuorodų populiarumo elektroninės skaičiuoklės:

http://www.kenkai.com/seo-tools-link-popularity.php

http://www.marketposition.com/linkpopularity.htm

http://www.linkpopularity.com/

4. TINKLAPIŲ OPTIMIZACIJA PAIEŠKOS SISTEMOMS

4.1. Paieškos trafiko svarba

Dominuojanti šiandienos globalaus tinklo tendencija yra ta, kad vartotojai dažnai paprastai tik "pažvelgia" į tinklapius, o ne „lanko" juos visa to žodžio prasme. Naudodamiesi paieškos sistemomis kaip sąsaja, vartotojai randa informaciją apie juos dominantį klausimą ir nesidomi pačių tinklapių turiniu.

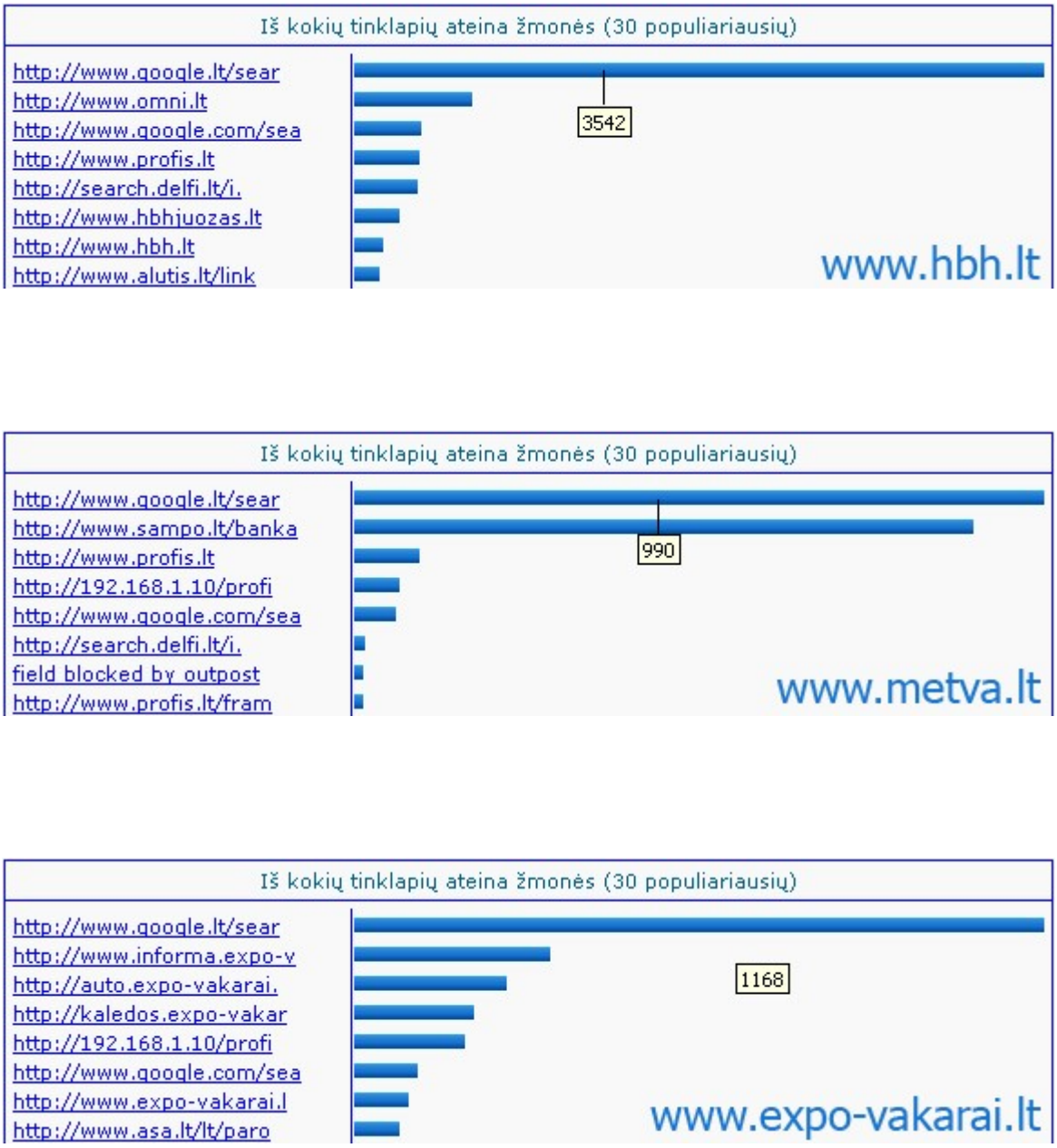
Pagal Jakobo Nilsono (Jacob Nielsen), žymiausio pasaulinio „usability" specialisto, tyrimus, - apklausus įvairias Interneto vartotojų reprezentacines grupes, į klausimą „Ką jie veikia Internete ir kokius tinklapius labiausiai vertina?" – visi atsakė skirtingai. Tačiau jų atsakymai turėjo vieną bendrą bruožą: kiekvienas vartotojas įvardijo paieškos sistemą tarp kitų 2-3 vertingiausių, jo nuomone, tinklapių [27].

Užuot ieškoję tinklapių ir įdėmiai juos nagrinėję, vartotojai paprasčiausiai ieško atsakymų į konkrečius klausimus. Tokiu būdu Internetas tapo resursų sąvadu, kurį vartotojai naudoja, kad surastų atskirus tinklapius pagal dominančią temą, nekreipdami dėmesio į tai, kokie tinklapiai jiems pateikiami. O paieškos sistemos tapo būdu surasti atsakymus. Jų užduotimi tapo ne resursų paieška, o atsakymų į vartotojų klausimus paieška.

4.2. Tinklapių optimizacija (SEO)

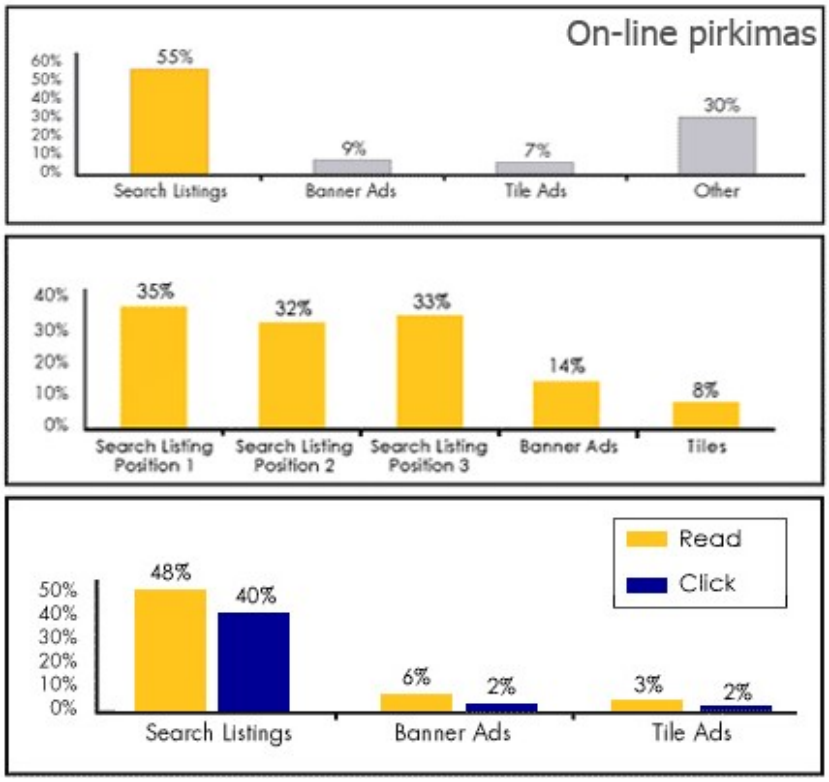
Paieškos optimizacija (SEO) - tai organizuota renginių sistema, nukreipta į tai, kad, veikdamas su bet kuria paieškos sistema, tinklapis turėtų aukštą reitingą [42]. Optimizacija - tai procesas, kurio efektyvumo negalima įvertinti iš karto. Pirmieji rezultatai pasirodo maždaug po mėnesio – tai apytikris laikas, kuris reikalingas, kad paieškos sistema pilnai indeksuotų tinklapį.

Kaip žinoma, pagrindinis lankytojų šaltinis šiandien – tai paieškos sistemos. Šito įrodymas – autoriaus 60 kompanijos Profis (www.profis.lt) tinklapių stebėjimas pusmečio laikotarpiu. Stebėjimų rezultatai atitinka oficialius tyrimus: **80-85%** lankytojų ateina į tinklapį iš paieškos sistemų. **95%** atvejų iš Google (15 pav.)



15 pav. ***Kompanijos ProfIS (www.profis.lt) duomenimis, 55 iš 60 tinklapių pirmoje vietoje yra Google, pirmaujanti 5-10 kartų***

Rinkdamas paieškos eilutėje savo užklausą, vartotojas gauna sąrašą tinklapių, atitinkančių užklausos tematiką. Tokių tinklapių gali būti 10, 100, 10.000. Paieškos sistemos pateikia paieškos rezultatus po 10-20 tinklapių viename puslapyje, be to, rūšiuoja tinklapius pagal releventiškumą. Peržiūrėjus analitinės kompanijos **NPD Group** 2001 m. sausio mėn. Statistiką (16 pav.), galima daryti išvadas – kodėl taip svarbu pakliūti į pirmąją dešimtį:



16 pav. ***Analitinės kompanijos NPD Group 2001 m. sausio mėn. statistika***

Paieškos optimizacija – procesas, kuris reikalauja reguliarių tyrimų, kad nebūtų pasimesta nuolatos besikeičiančiame paieškos sistemų konkurencijos pasaulyje. Paieškos sistemų darbo algoritmai – tai visada griežtai saugojamos paslaptys, kadangi bet koks informacijos nutekėjimas ves prie paieškos rezultatų kokybės pablogėjimo.

Autoriaus nuomone, tarp Lietuvos paieškos sistemų galima pažymėti tik **Search.lt** ir **Delfi.lt**, kurių darbą dar reikia tobulinti. Todėl optimizuojant tinklapį reiktų orientuotis į tarptautines paieškos sistemas, tarp kurių yra ir Google. Tačiau, atsižvelgiant į Google ir kitų paieškos sistemų oficialių direktyvų ir komentarų trūkumą, autoriaus darbo tyrimai remiasi įvairių prielaidų patikrinimu, kitų tyrinėtojų paskelbta medžiaga, savo patirtimi ir paieškos sistemų funkcionavimo, vystymosi tendencijų žiniomis.

4.3. SEO trūkumai ir problemos

Užklausų konkurencingumas

Paieškos optimizacija dažniausiai veikia tik konkurencingoms užklausoms. Konkurencingiausios yra šių temų užklausos: „kazino“, „draudimas“, „sveikata“, „viešbučiai“ ir t.t. Per daug žmonių siekia aukštų pozicijų šia tema ir tai sukuria daug sunkumų optimizatoriui.

Pritaikymo galimybės

Ne visi tinklapiai tinka SEO. Daug tinklapių paprastai neturi pakankamai teksto, o jo padidinimas gali pakenkti dizainui arba tinklapio charakteriui. Be to, kai kurie tinklapių kūrėjai nenori keisti tinklapio tekstų tik dėl optimizacijos.

Prisirišimas prie optimizatoriaus

Optimizuotame tinklapyje teksto negalima keisti, nepakenkiant brangiam optimizavimo darbui, nes tai gali sumažinti reitingą. Išeitis – pakartotinas kreipimasis į tą patį optimizatorių, kad po teksto pakeitimo jis vėl atliktų reikalingą darbą.

Rezultatų pateikimo pasikeitimai.

Jeigu tinklapis buvo tinkamai optimizuotas ir yra pirmoje dešimtyje pagal tikslinę užklausą, tai optimizacija buvo verta jai skirtų lėšų. Bet kas vyksta, jei kas nors iš konkurentų irgi nusprendžia profesionaliai optimizuoti savo tinklapį pagal tą pačią užklausą? Rezultatų pateikimo pasikeitimai ir galimas pirmų vietų praradimas, iškritimas iš dešimtuko – o tai reiškia naujos reoptimizacijos būtinybę.

4.4. Praktiniai patarimai optimizacijai

Raktinių frazių nustatymas

Gerų paieškos žodžių, kurie didina tinklapio lankomumą, paieška – tai pirminis uždavinys vykdant paieškos optimizaciją. Rasti 100 bendrų užklausų yra lengviau, negu sukurti trumpą relevantiškiausių ir efektyviausių žodžių sąrašą. Didelę reikšmę turi teisingos raktinių žodžių proporcijos.

Tinklapio struktūros teisingumas

Jeigu paieškos robotai nesugebės perskaityti tinklapio struktūros, tai jokia optimizacija nepadės tinklapiui užimti gerų vietų reitinguose.

Tinklapio registravimas

Tinklapį reikia registruoti ne tik Google ir DMOZ’e, bet būtinai ir teminiuose kataloguose.

Partnerių apsikeisti nuorodomis paieška

Gerų, panašios tematikos kokybiškų tinklapių, norinčių patalpinti pas save nuorodą į reklamuojamą tinklapį, paieška – tai nelengvas uždavinys. Jo sprendimui reikalingas sugebėjimas vesti pokalbius ir gausybė laiko, kad būtų sukurta gerų relevantiškų partnerių bazė.

Semantinė optimizacija

Įdomus ir aktualus tinklapių turinys pritraukia lankytojus. Svarbu talpinti puslapiuose kuo daugiau nuorodų „žiūrėkite taip pat čia...“ tipo. Jos veda prie tekstų ar paslaugų, turinčių ryšį su užklausos tema. Tokios kontekstinės nuorodos priverčia susidomėjusius lankytojus „eiti gilyn“ – o būtent tokie lankytojai gali tapti potencialiais klientais.

4.5. Reklama

4.5.1. Kontekstinės reklamos, banerių ir nuorodų palyginimas

Kontekstinė reklama – tai tekstinis skelbimas su nuoroda į reklamuojamą tinklapį, atsirandantis paieškos sistemoje arba teminiame tinklapyje, kai pateikiami tam tikros užklausos, atitinkančios reklamą, rezultatai arba tiesiog peržiūrimas teminis tinklapis.

Skirtingai nuo banerinės reklamos (reklama naudojant banerius), kontekstinė reklama siūlo lankytojui prekę arba paslaugą būtent tuo momentu, kai jos jam labiausiai reikia (t.y kada jis ieško kažko su panašia tematika). Teisingai pasirenkant užklausas ir kokybiškai sudarant reklamos tekstą, galima pasiekti labai gerų rezultatų, nes kontekstinė reklama pritraukia tik tikslinę auditoriją.

Palyginus kontekstinę reklama su nuoroda, nuoroda turi žymiai didesnę vertę. Pagal psichologų atliktus lankytojų elgsenos tyrimus, tiktai apie 5% lankytojų iš visos paieškos sistemų auditorijos žiūri kontekstinės reklamos skelbimus [34]. Neoptimizuojant tinklapio ir naudojant tik kontekstinės reklamos galimybes, prarandama didelė auditorijos dalis. Todėl, optimizuojant tinklapį, kontekstinė reklama gali atlikti tik pagalbinio įrankio vaidmenį.

4.5.2. Antireklama

Google reitingo skaičiavimų ypatumai leidžia padaryti taip, kad tinklapis, į kurį yra nukreipta antireklaminė kompanija, bus pirmose pozicijose pagal tam tikrus antireklaminius raktinius žodžius, kurių tinklapyje-aukoje iš vis nėra! Antireklamos rezultatą galima stebėti 17 pav.

17 pav. ***Antireklamos Google pavyzdys (2005-05-16)***

Kompanijos **InteraX** (www.intera.lt)**,** užsiimančios tinklapių kūrimu, tinklapyje iš viso nėra žodžių **"tikri luzeriai",** bet jų tinklapis tvirtai yra pirmoje vietoje. Antroje pozicijoje stovi kompanijos **ProfIS** tinklapis, kuriame ir patalpintas straipsnis apie antireklamą. Toks antireklamos metodas vadinasi "**Google Bomb"** [43]. Jis pagrįstas tuo, kad autoritetingas šaltinis patalpina savyje relevantišką kontekstinę nuorodą į tinklapį-auką su antireklaminiais žodžiais.

4.6. Technologijos, kurios trukdo efektyviam paieškos sistemų darbui

Paieškos sistemos mėgsta paprastumą, todėl kai kurios technologijos, naudojamos kuriant tinklapius,tinka žmogui, tačiau neperskaitomos paieškos robotui. Štai tos technologijos:

Paveikslėliai ir grafiniai nuorodų žemėlapiai (areas).

Dažnai dizaineriai kuria visiškai grafišką tinklapį, o kadangi robotas nemato paveikslėlių, tai toks tinklapis paieškos sistemoms yra negyvas.

Navigacijos meniu.

Iššokantis JavaScript meniu palengvina vartotojui navigaciją, tačiau kadangi robotas keliauja nuorodomis, bet neskaito jokių skriptų, tai jam tos „nuorodos" neegzistuoja – ir atitinkamai jis negali indeksuoti gilesnių puslapių.

Dinamiški puslapiai.

Tokie puslapiai kuriami iš karto, kai vartotojas bando juos peržiūrėti, ir jų URL gali turėti daug robotui nesuprantamų simbolių, o tai trukdo robotui normaliai keliauti nuorodomis.

Freimai.

Daugelis paieškos sistemų iš viso atsisakė dirbti su freimais, todėl freimų buvimas tinklapyje nepageidaujamas.

4.7. Raktiniai žodžiai (keywods)

Raktiniai žodžiai – tai žodžiai, kurie yra ir dokumento tekste, ir numanomoje vartotojo užklausoje, kuri sutampa su dokumento tema [16]. **Raktinių frazių** vartojimas žymiai efektyvesnis už atskirų raktinių žodžių vartojimą. Raktinės frazės kūrimas – vienas svarbiausių optimizacijos aspektų. Daugelis SEO ekspertų mano, kad pagrindiniai raktiniai žodžiai turi būti pirmuose 2-3 teksto sakiniuose. Taip pat reikia atsižvelgti į tai, kad jeigu reitinge pagal tikslinę užklausą konkurentų tinklapiai kokybiški ir turi PR5-PR6, tai galima tvirtinti, kad tinklapiui reikės mažiausiai tokio pat PR kovai už pasirinktus raktinius žodžius.

Praktinių eksperimentų metu buvo sukurta raktinių žodžių žodyno kūrimo metodika: Pirmiausia sudaromas pagrindinių žodžių sąrašas. Į jį privalo įeiti:

- Prekės ar paslaugos pavadinimo variantai, sinonimai, vertimai, transliteracija.
- Slengas, profesiniai ir sutrumpinti pavadinimų variantai.
- Prekės ar paslaugos sudėtinių dalių pavadinimų variantai.
- Lankytojo uždaviniai ir problemos, kurias šios prekės ar paslaugos išsprendžia (ir pagal kuriuos jų galima ieškoti).
- Prekės ar paslaugos kokybės, patrauklumo savybės (patikimas, greitas, profesionalus).

Toliau reikia praplėsti sąrašą su statistikos duomenų pagalba. Egzistuoja tarnybos, kurios leidžia patikrinti raktinių žodžių populiarumą:

KeywordSandbox: https://adwords.google.com/select/KeywordSandbox

WordTracker: http://www.wordtracker.com/

Overture: http://inventory.overture.com/d/searchinventory/suggestion/

GoogleSets: http://labs1.google.com/sets.

Daug įdomios informacijos apie vėlyviausią statistiką ir užklausų archyvo duomenis galima rasti **Google Zeitgeist** (18 pav.) puslapyje: http://www.google.lt/press/zeitgeist.html

Zeitgeist This Week

Top 10 Gaining Queries Week Ending May 9, 2005	Top 10 Declining Queries Week Ending May 9, 2005
1. mother's day	1. katie holmes
2. kentucky derby	2. ivory-billed woodpecker
3. cinco de mayo	3. jennifer wilbanks
4. orlando bloom	4. zooey deschanel
5. paula abdul	5. wynn las vegas
6. xbox 360	6. nashville star
7. kingdom of heaven	7. airbus
8. jennifer garner	8. constantine maroulis
9. angelina jolie	9. lorraine kelly
10. kenny chesney	10. lynndie england

18 pav. ***"Google Zeitgeist: Search patterns, trends, and surprises according to Google"***

Dar vienas metodas – raktinių žodžių panaudojimas pagal aktualias užklausas. Pvz.: 2001 m. kompanija **WordTracker** paruošė specialų savo kasmetinio biuletenio "**Wordtracker keyword report"** leidinį. Jame buvo 300 pačių populiariausių užklausų per paskutines 24 val. (šie duomenys buvo gauti iš populiariausių paieškos sistemų). Jame atsispindi tų metų Amerikos įvykiai:

1 vieta - 5826 osama bin laden
2 vieta - 4332 nostradamus predictions
3 vieta - 3902 Osama bin Laden

Beje, čia galima pastebėti dar vieną įdomią tendenciją – žmonės rašo užklausas vartodami tik mažąsias raidės. Analogišką vaizdą galima buvo stebėti 2004 m. balandžio mėn. (Romos popiežiaus Jono Pauliaus Antrojo mirtis).

Perkami raktiniai žodžiai.

Daugumoj sistemų raktinius žodžius tiesiog galima pirkti. Kai kuriose tai garantuoja aukštą reitingą, o kitose (pvz. Google) – šalia natūraliųjų paieškos rezultatų rodomas reklaminis skelbimas, atitinkantis raktinę frazę arba relevantišką tekstą, neįtakojant jų.

Relevantiškumas ir raktinių žodžių svoriai

Priklausomai nuo konkurencijos lygio ir optimizacijos tikslų, raktinių žodžių kiekis tekste gali būti nuo 20 iki 30. Kuo mažiau raktinių žodžių dokumente, tuo didesnis gali būti jų dažnis ir tuo relevantiškesni jie turi būti. Aukšto relevantiškumo dokumentai orientuojasi į 2-3 raktinius žodžius ir kiekvienas iš jų turi užimti iki 5% viso teksto. Bendra rekomendacija - naudoti raktinius žodžius tekste su 3-7% proporcija, ir testai patvirtina, kad tai yra optimalus variantas. Jeigu raktinių žodžių yra mažiau negu 3%, jų svoris bus mažesnis.

Daugelis paieškos sistemų teigia, kad aukštas raktinių žodžių dažnis laikomas spamu. Autoriaus stebėjimai aiškina šį teiginį. Tarkime, kad jei puslapyje yra 3 žodžiai ir jie visi raktiniai, tada jų tankis bus 100%. Bet tokie puslapiai iš tikrųjų labai sėkmingai indeksuojami. Reiškia, yra kažkokios **slenkstinės reikšmės**, kurias pasiekus, tolesnis dažnio didinimas bus laikomas spamu ir neįtakos relevantiškumo.

Svarbu pažymėti, kad kartais tinklapis, paieškos rezultatuose esantis №5, pritraukdavo daugiau lankytojų, negu №1 su netinkamu aprašymu. Pvz. vietoj paprasto aprašymo "tooth extraction" (dantų traukimas), užrašas "painless tooth extraction" (beskausmis dantų traukimas) akivaizdžiai pritrauks daugiau susidomėjusių lankytojų [34].

4.8. Užklausos ir kompleksinė optimizacija

"Viskas, ko mums reikia, jau kažkur egzistuoja. Reikia tik tinkamai suformuluoti užklausą "

J. Nielsen

Didžioji lankytojų dalis, formuodami paieškos kriterijus, nenaudoja jokių ypatingų intelektualinių pastangų. Apie 80% visų užklausų susideda iš vieno žodžio. Užklausų su nors vienu loginiu operatoriumi procentas yra labai žemas (ne daugiau nei 20%). Jeigu operatoriai ir naudojami, tai daugumoj tai yra AND ir OR. Kontekstinio artumo ir loginio neigimo operatorių dalis yra tik 1-2%. Bet būtent sudėtingos užklausos lemia laiko, kurį lankytojas praleidžia Internete, panaudojimo efektyvumą [23].

Egzistuoja 2 tinklapių optimizacijos būdai:

- Greitas - tik pagal užklausas (raktinius žodžius).
- Kompleksinis (nuorodų, kodo, turinio ir t.t. optimizacija).

Kompleksinė optimizacija – tai visa darbų su tinklapiu eilė:

- Tinklapio semantinio branduolio kūrimas.
- Konkurencijos pagal užklausas ir konkurentų padėties analizė. Optimizavimo strategijos kūrimas ir lankomumo prognozė.
- Tinklapio kodo ir struktūros analizė ir optimizacija skirtingoms paieškos sistemoms.
- Turinio plėtimas ir tobulinimas, lankytojų poreikių ir motyvų analizė.
- Apsikeitimas nuorodomis su kokybiškais tematiniais tinklapiais.

Kiekvienas tų dviejų būdų turi savo privalumų/trūkumų ir savo rizikos laipsnį. Tačiau, net ir optimizuojant tinklapį pagal „n“ užklausas, į jį ateis lankytojai pagal „n+m“ užklausas. T.y., bendras trafikas iš paieškos mašinų bus didesnis, negu tik konkrečių užklausų trafikas.

Kaip pavyzdį galim panagrinėti kompanijos ProfIS tinklapį, kuriame 2005 m. sausio mėn. buvo atlikta kompleksinė optimizacija. Lankomumo rezultatus galima stebėti 19 pav.:

19 pav. ***ProfIS tinklapio lankomumas iki kompleksinės optimizacijos ir po jos (http://www.profis.lt/statistika)***

4.9. SEO metodai ir strategijos

Kaip paieškos sistemos besistengtų paaiškinti tinklapių savininkams, kad vieta paieškos rezultatuose niekam negarantuota, tačiau nauji ir sėkmingi optimizavimo metodai atrandami iki šiol. Paieškos sistemų netobulumas leidžia daryti tokius triukus, kaip visų pozicijų pirmame rezultatų puslapyje užėmimas (20 pav.):

20 pav. ***Kompanijos ProfIS tinklapiai paieškos sistemoje Delfi.lt užima 7 iš 10 pirmo rezultatų puslapio pozicijų pagal frazę "tinklapiu kurimas" (2005-05-16)***

4.9.1. Legali optimizacija

Legali optimizacija – papildymai bei tinklapio turinio ir struktūros klaidų taisymas, kad būtų galima užimti **užtarnautą** aukštą vietą paieškos sistemų reitinguose. Paprastai tai labai daug darbo reikalaujanti ir lėta procedūra (tinklapio puslapių kodo pakeitimai, naujų puslapių pridėjimas, tinklapio vidinių nuorodų struktūros keitimas, apsikeitimas nuorodomis su kitais teminiais tinklapiais, registracija kataloguose ir paieškos sistemose ir t.t.)

4.9.2. Nelegali optimizacija ("Spamdexing")

Nelegali optimizacija (spamdexing) – tai uždraustų metodų panaudojimas ir paieškos sistemų apgavystė, norint užimti neužtarnautą aukštą vietą paieškos sistemų reitinguose. Tai faktiškai yra spamas, nuo kurio kenčia ir eilinis vartotojas, kuris nerado to, ko ieškojo, - ir paieškos sistema, kurioje mažėja paieškos kokybė. O iš tikrųjų daugiausia kenkiama pačiam tinklapiui, nes anksčiau ar vėliau jis pateks į paieškos sistemų uždraustų tinklapių sąrašą [32].

Perėjimai su nukreipimu (Doorways with redirect)

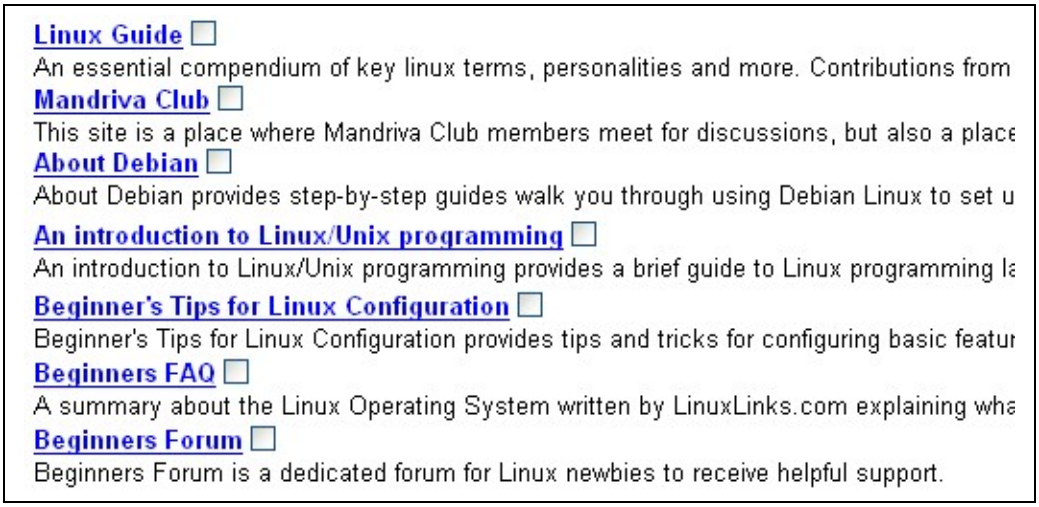
Tai specialūs puslapiai su beprasmiu vartotojui tekstu, bet užpildyti raktiniais žodžiais, reikalingais paieškos robotams. Taip pat šiame puslapyje kuriamas nukreipimas į pagrindinį šakninį tinklapio puslapį. Toks puslapis reklamuojamas ir patenka į aukštą reitingo vietą. Tokiu būdu lankytojas, paspaudęs nuorodą, patenka į perėjimo puslapį ir automatiškai nukreipiamas į pagrindinį ir net nepastebi to. Rasti tokius tinklapius Internete labai lengva. Užtenka tik duoti užklausą „**seksas referatai porno**“. Tokia technologija labiausiai paplitusi tarp adult-tematikos tinklapių, bet dažnai naudojama ir kitur.

Maskavimas (Cloaking)

Šis terminas sujungia daug metodų, leidžiančių pateikti lankytojams ir paieškos robotams skirtingas tinklapio versijas. Techniškai tai realizuojama taip: specialūs skriptai sulygina lankytojo IP-adresą su paieškos robotų adresų DB duomenimis. Jeigu tai yra paprastas lankytojas, tai jam pateikiamas normalus tinklapis, o jeigu robotas – tai jam pateikiamas specialiai optimizuotas tinklapis, dažnai net neperskaitomas žmogui. Norint atpažinti, ar tinklapyje naudojamas toks metodas, užtenka tik palyginti esamo tinklapio kodą su indeksuotu tinklapio kodu, kuris išsaugotas paieškos sistemos atmintyje.

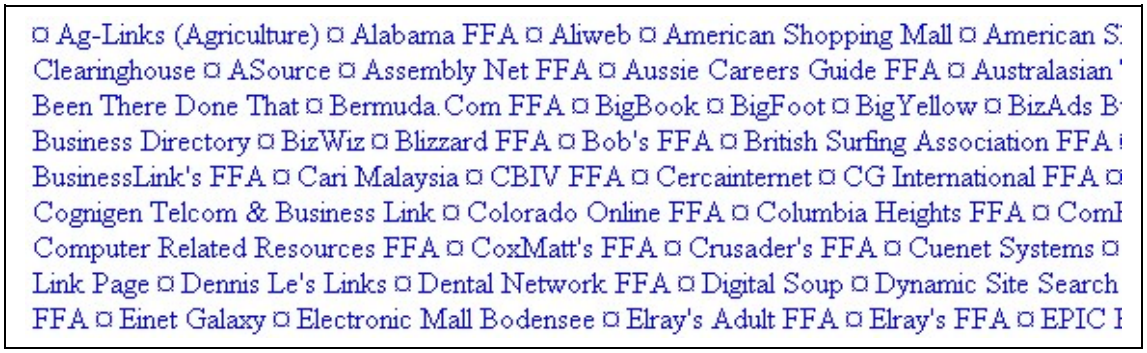
Nuorodų fermos (Link Farms) ir Free-for-All (FFA) katalogai

Nuorodų fermos - tinklapiai, specialiai skirti talpinti tik tam tikras nuorodas (tinklapių populiarumui padidinti). Teisingai įvertinti, ar tai leistinas tinklapis, ar nuorodų ferma, labai sunku, nes sunku objektyviai įvertinti nuorodos talpinimo teisėtumą. Pvz. ryškiai atskirtus nuorodų blokus (katalogo pavidalu) paieškos sistemos toleruoja (21 pav.):



21 pav. ***Ryškiai atskirti nuorodų blokai***

O tinklapių savininkai, kurie talpina pas save betvarkį nuorodų rinkinį, gali patekti į uždraustų tinklapių sąrašą. Tokie tinklapiai paprastai generuojami automatiškai (22 pav.)



22 pav. ***Nuorodų ferma***

Dar viena fermos rūšis – **FFA (Free For All) katalogas.** FFA kataloguose vyksta nuolatinė nuorodų rotacija, ir todėl ten reikia registruotis iš naujo ir iš naujo – o tai didelės laiko sąnaudos ir nereikalingas darbas. Be to, daugelis FFA tinklapių tiesiog skirti tik el. paštų, į kurios galima siųsti spamą, rinkimui. Beveik visi paieškos sistemų nuorodų iš tokių fermų neįskaito.

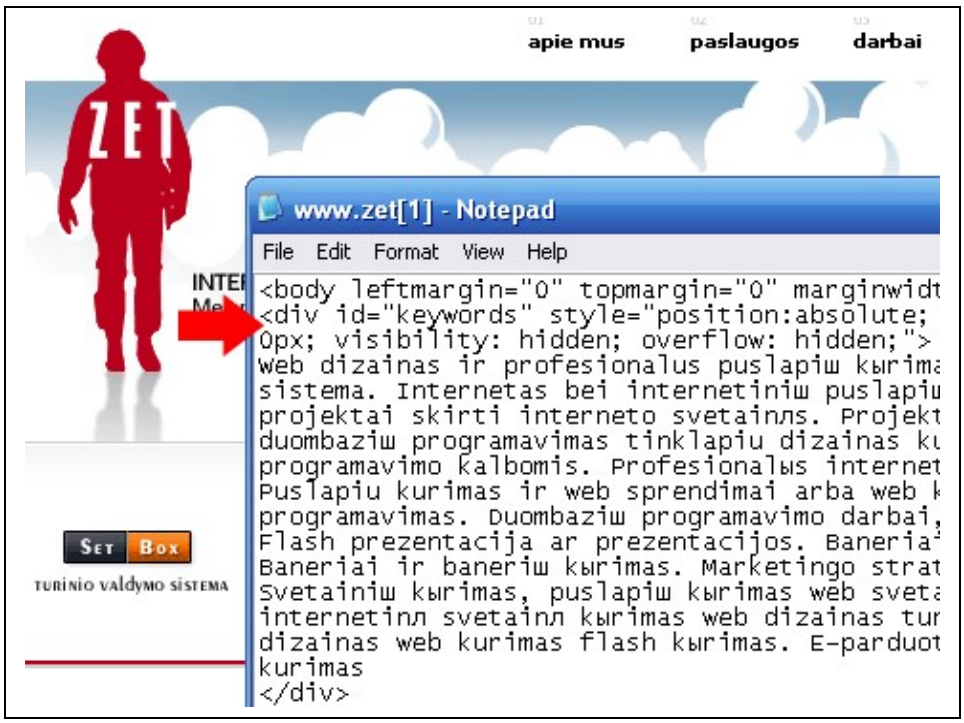
Nematomas ir paslėptas tekstas ar nuorodos

Tai populiariausių raktinių žodžių užrašymas tinklapio fono spalva. Tie žodžiai paprastai neturi jokio ryšio su tinklapio tema. Neseniai tokia technologija dirbo gerai, bet dabar, kai paieškos sistemų algoritmai nuolat tobulėja, tekstiniai kriterijai jau yra nelabai svarbūs. Be to, paspaudęs «Ctrl+A», lankytojas gali pažymėti visą tekstą ir pamatyti nematomą tekstą.

Egzistuoja keletas nematomo ar paslėpto teksto įterpimo į puslapį būdų:

- Naudoti spalvą, nežymiai tesiskiriančią nuo fono spalvos (pvz. #ffffff и #feffef).
- Paslėpti tekstą į „konteinerį“ ( ) ir išnešti už puslapio matomumo ribos su atributo „style“ pagalba. Pvz.: style=”position: absolute, top:-1000px;”. Tokį metodą galima papildomai apsaugoti, jei stilius bus taikomas per CSS arba JavaScript.
- Patalpinti tekstą į tegus <noframes> arba <style> - jie nematomi lankytojui.

Praktika rodo, kad tokio rizikingo metodo panaudojimas didelės naudos neatneša ir tik kartais gali duoti trumpalaikį reitingo prieaugį.



23 pav. ***Kompanija „Zenito Technologijos“ (www.zet.lt) naudoja triuką su paslėptais raktiniais žodžiais (2005-05-16)***

Raktinių žodžių pertekliai (Keyword stuffing)

Raktinų žodžių dažnio didinimas šiais laikais jau nelaikomas gera idėja, nes visos paieškos sistemos turi tankio slenkstinę reikšmę, už kurios visi raktiniai žodžiai laikomi spamu. Tačiau jeigu puslapiui trūksta raktinių žodžių, jų sąrašą galima įterpti paslėpto teksto pavidalu (23 pav.). Tokiu atveju reikia vengti paprasto žodžių užrašymo per kablelį, nes paieškos sistema laiko tai įtartinu nenatūralaus teksto bloku ir gali pažymėti kaip spamą.

Veidrodiniai tinklapiai (Mirror websites)

Tai keletas skirtingų URL adresų, kurie nukreipti į tą patį tinklapį. Veidrodžių panaudojimo paieškos sistemos netoleruoja ir filtruoja juos iš indekso, palikdamos tik pagrindinį tinklapį. Išeitis iš tokios padėties yra. Tai tinklapio keitimas priklausomai nuo URL, tačiau vizualiai lankytojai gali to nematyti. Pakeitimai atliekami vidinių nuorodų sistemoje, nuorodų tekste, raktiniuose žodžiuose, žodžiuose metateguose ir <title> tege bei paties teksto struktūroje. Galimi ir minimalūs pakeitimai dizaine. Tokiu būdu robotas, ateidamas į veidrodinį tinklapį, laikys jį atskiru tinklapiu, kuris turi teisę dalyvauti reitinge.

Typo-spam

Tinklapius su labai panašiais į žinomų originalių saitų adresais, pvz.: www.yahhoo.com, www.gogle.com, www.googlee.com paieškos sistemos laiko spamerių tinklapiais. Laikoma, kad lankytojas patenka ten netyčia, dėl rikto URL adrese.

Jeigu paieškos robotas pastebės tinklapį, naudojantį aukščiau išvardytas nelegalias technologijas, tinklapis greičiausiai bus pašalintas iš reitingo visam laikui. Tokiu atveju galima vienintelė išeitis: nupirkti naują domeno vardą ir sukurti visiškai naują tinklapį.

Patikrinti tinklapį, ar jis neturi paieškos spamo, galima čia:
http://tool.motoricerca.info/spam-detector/

Paskųsti pažeidėją galima čia:
Google: http://www.google.com/contact/spamreport.html
Altavista, Yahoo: http://help.yahoo.com/help/us/ysearch/deletions/index.html
HotBot, Lycos : http://help.lycos.com/watchdog/watchdog_form.asp

4.10. Apsikeitimo nuorodomis strategija

Kadangi dauguma paieškos sistemų vertina nuorodų kiekį ir jų kokybę, tinklapių optimizacija numato nuolatinį nuorodų į optimizuojamą tinklapį kiekio augimą. Nuorodų kokybė nustatoma pagal tai, ar rimti tinklapiai nurodo į optimizuojamą tinklapį ir kiek jų tematika sutampa su optimizuojamo tinklapio tematika [42].

Dažnai tinklapių savininkai peržiūri Internetą, ieškodami tinklapių, kurie nori apsikeisti nuorodomis. Bet ką jie abudu paprastai daro, kai tik suranda apsikeitimo partnerį? Jie stengiasi paslėpti partnerio nuorodą kuo giliau (kad paieškos robotas ar lankytojai jos nerastų), laukdami, kad anas partneris patalpins jų nuorodą tinkamai. Kadangi tokiu būdu nelaimi niekas, tai patikimų partnerių reikia ieškoti atidžiai.

Jeigu tinklapio PR lygus 6 ar daugiau – jis yra tinkamas apsikeitimo kandidatas.

Tvarkant tinklapį kartais reikia išsaugoti visą PR jo viduje. Tai galima padaryti keliais būdais:

- Nurodyti į išorę tik iš puslapio su žemiausiu PR.
- Bet kuri išorinė nuoroda, apiforminta su JavaScript, bus normaliai suprantama vartotojui ir naršyklei, bet nebus įskaitoma paieškos sistemai skaičiuojant PR.

Naudinga talpinti nuorodas forumuose ir laikraščių skelbimuose bei atsiliepimuose apie tinklapį (kur tai leidžiama). Jeigu koks nors puslapis turi nuorodą į kitus tinklapius, svarbu įsitikinti, kad paieškos sistemos tų tinklapių nenubaudė, nes nuoroda į juos gali atvesti prie to, kad jūsų tinklapis irgi bus nubaustas [10]. Taip pat svarbu periodiškai tikrinti, ar tinklapiai, į kuriuos yra nuorodos, nepateko į juodąjį paieškos sistemų sąrašą (pagal Google paskutinio patento duomenis "*Outbound links to low trust or affiliate websites may be an indicator of low quality (section 0089)*" [1]). Taip pat svarbu atsižvelgti į tai, kad per daug greitas nuorodų kiekio augimas irgi gali būti laikomas spamo požymiu. Net ir viena nuoroda, vedanti į niekur (**broken link**), gali smarkiai pažeminti paieškos rezultatų pozicijas.

4.11. Failas „robots.txt"

Failas „robots.txt" skirtas nurodyti paieškos robotams, kokie tinklapio puslapiai ar katalogai negali būti indeksuojami. Liujiso Monje (Louis Monier iš Altavista) nuomone, tik 5% visų tinklapių šiuo metu turi netuščią „robots.txt" failą. Tačiau failas „robots.txt" – tai pirmas dalykas, į kurį atsižvelgia voras, indeksuodamas tinklapį. Todėl svarbu turėti net tuščią failą. Yra nuomonių, kad šis failas įtakoja indeksaciją (tuščias failas reiškia, kad tinklapis neturi ko slėpti – o tai pagal SEO ekspertų prielaidas gali teigiamai įtakoti tinklapių vertinimą) [17].

Failas „robots.txt" naudoja tokią sintaksę:

User-agent: {SpiderName}
Disallow: {Filename}

Išsamus vorų ir jų duomenų sąrašas patiektas **2 Priede.**

4.12. Domeno vardas

Teisingai pasirinktas domeno vardas labai svarbus viso tinklapio projekto pasisekimui. Dažnai lankytojai patenka į tinklapį tiesiog atspėdami jo pavadinimą. Taip pat Google, skaičiuodamas reitingus, atsižvelgia į:

- Domeno amžių.
- Domeno savininko adresą (didina svorį vietinėms užklausoms).
- Informaciją apie domeną (jeigu ji dažnai kinta – įtariama, kad tai spamas).

4.13. Tinklapio atnaujinimas

Pagal Google paskutinį patentą [1], ji atskiria nežymius atnaujinimus nuo reikšmingų ir reaguoja tik į reikšmingus. Kad atpažintų naujumą, Google fiksuoja tokius faktus:

- Visų tinklapio puslapių atnaujinimo dažnumas.
- Atnaujinimo lygis (reikšmingas ar ne).
- Raktinių žodžių dažnio ir tankio pasikeitimai.
- Naujų puslapių su nuorodomis į šį puslapį kiekis.
- Nuorodų turinio kitimas.
- Naujų nuorodų į spamerių / mažos kokybės tinklapių kiekis.

4.14. Metategai ir tinklapio pavadinimas

Daugelis neigia metategų įtaką indeksavimui, tačiau autoriaus tyrimai patvirtino jų įtaką. Tos pat nuomonės laikosi ir autoritetingas optimizatorių tinklapis www.bruceclay.com [9], o MSN net teigia, kad jie tikrai įtakoja reitingą. Egzistuoja daug skirtingų metategų, bet realią naudą turi tik 2 (o kai kurie, pvz. **meta-redirect**, atvirkščiai – žalingi tinklapiui).

TITLE

Tai pats reikšmingiausias tinklapio tegas. Jame turi būti patalpinti svarbiausi raktiniai žodžiai ir frazės. Kompanijos ar tinklapio pavadinimą siūloma talpinti šio tego pabaigoje, o ne pradžioje, kaip daugelis daro. Jo ilgis neturi viršyti 10 žodžių arba 90 simbolių įskaitant tarpus.

KEYWORDS

Žodžiai, esantys šiame metatege, paieškos sistemų vertinami kaip rekomendacijos. Viena iš labiausiai paplitusių klaidų – neteisingas šito tego pildymas. Jeigu visi žodžiai jame parašyti per tarpą, o ne per kablelį, jie bus laikomi viena fraze, ir todėl iš šito tego nebus jokios naudos. Kuo daugiau raktinių žodžių naudojama, tuo ne toks reikšmingas bus kiekvienas jų. Bet kuris raktinis žodis, paminėtas tinklapyje, turi būti įrašytas į šitą tegą, ir atvirkščiai, - kitaip tai gresia reitingo sumažėjimu kaip neatitikimas temai. Raktinių žodžių užrašymas didžiosiomis ir mažosiomis raidėmis, užrašymo variantai ir įprastos rašybos klaidos yra absoliučiai teisėta ir efektyvu (Pvz.: «įranga» ir «iranga», nes žmonės paprasčiausiai gali pamiršti perjungti kalbą).

DESCRIPTION

Šis metategas naudojamas paieškos sistemų kaip trumpa anotacija šalia nuorodos atsakant į užklausą. Būtent pagal šio tego turinį paieškos sistemos vartotojas įvertins tinklapio tinkamumą. Jeigu šis metategas nenaudojamas, tai kaip aprašas tinka teksto pirma eilutė arba ištrauka iš teksto su rastu raktiniu žodžiu. Nerekomenduojama kopijuoti į šį metategą tego <title> turinį.

4.15. Paveikslėliai

Tipiška klaida yra ta, kad paveikslėlio failas dažnai pavadinamas skaičiais arba žodžiais, kurie neturi jokio ryšio su paveikslėliu ar raktiniais žodžiais. Pvz. jei raktinis žodis yra «pienas», o paveikslėlį pavadina „aaa.gif“. Be to, paveikslėlio tegas <img> turi turėti atributą ALT. Rekomenduojama įrašyti į ALT atributą paveikslėlio pavadinimą arba trumpą aprašą (10-40 simbolių) ypač tada, kai jame yra raktinių žodžių. Bet nereikia paversti ALT atributo metatego KEYWORDS kopija. Tas pats tinka ir nuorodos tego <a> atributui TITLE. Reikia pažymėti, kad Google įskaito ALT atributo turinį tiktai tiems paveikslėliams, kurie yra nuorodos.

4.16. Autoriaus pasiūlytas „pakartojimo metodas“

Metodo esmė – vienų ir tų pačių raktinių žodžių vartojimas ir pakartojimas 8-iose vietose:

<!DOCTYPE …>
<HTML>
<HEAD>
<TITLE>**1. Raktiniai žodžiai**</TITLE>
<META Name="description" Content="**2. Žodžiai iš TITLE ir dar keletas žodžių, pageidautina - skirtingose vietose. Ilgis – iki 180 simbolių**">
<META Name="keywords" Content="**3. Raktiniai žodžiai + dar apie 20 žodžių**">
</HEAD>
<BODY>
<H1>**4. Raktiniai žodžiai iš TITLE**</H1>
5. Puslapio tekstas (pageidautina - patys pirmieji teksto žodžiai) turi turėti visus raktinius žodžius, išvardytus aukščiau.
</BODY>
</HTML>

6. Failo pavadinimas turi turėti vieną arba du raktinius žodžius:
Pvz.: **"keyword1-keyword2.html"** („-“paieškos sistemoms reiškia tarpą).

7. Nuorodų tekste turi būti raktiniai žodžiai. Nuorodos turi nurodyti į puslapius su pavadinimais, turinčiais raktinius žodžius:

<A HREF="http://www.keyword.lt/keyword.html">kewyword1-keyword2</A>

8. Domeno vardas ir subdomenų vardai turi turėti reikšmingiausius raktinius žodžius:

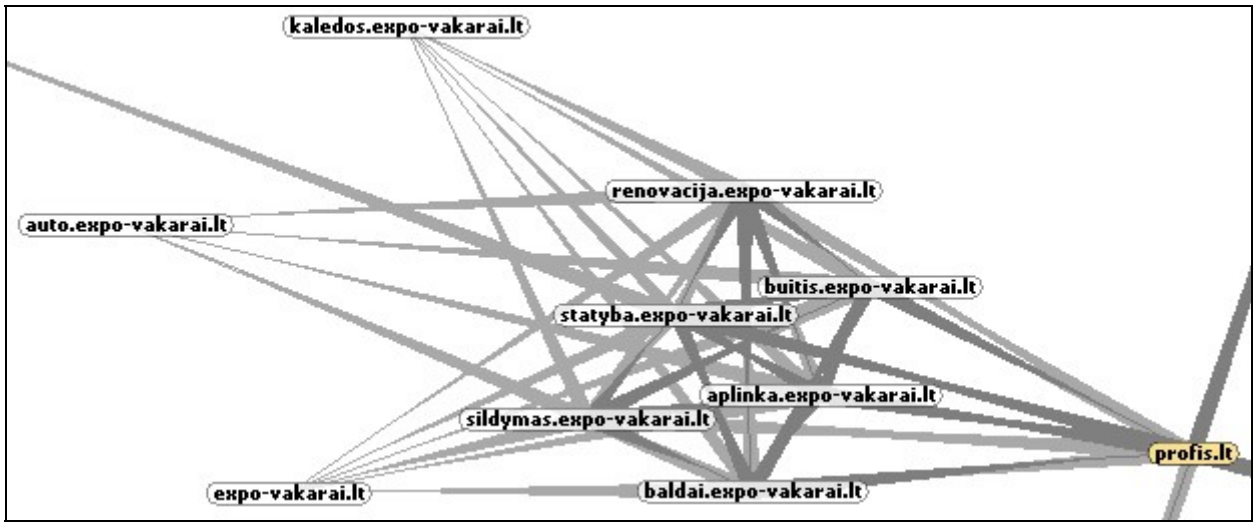
http://www.keyword.lt или **http://www.keyword1.keyword2.lt**

Taikant šį metodą, privalu atsiminti, kad:

- Reikšmingus žodžius reikia išskirti tegais ****, **<b>**, **<u>**, ****, o paantraštes – tegais **<h3>-<h6>.** Tegas **<h1>** turi didžiausią svorį ir turi būti vartojamas kaip puslapio antraštė.
- Negalima vartoti vienodų žodžių šalia daug kartų kiekvienoje iš 8 vietų (tai laikoma spamu). Tačiau vienodų žodžių pakartojimas įmanomas, jeigu jie yra toli vienas nuo kito.
- Puslapiai ir subdomenai turi būti surišti vidinėmis nuorodomis. Būtinas nuorodos į tinklapio žemėlapį buvimas bet kuriame tinklapio puslapyje.
- Teksto dydis turi būti ne mažesnis nei 500 žodžių.
- Indeksuojamo failo dydis ne turi būti daugiau negu 100Kb (Google, MSN)

4.17. Subdomenai ir tinklapių klasteris

Jeigu egzistuoja galimybė tinklapiui kurti teminius subdomenus – ja reikia būtinai pasinaudoti. Tinkamai sudarius vidinių nuorodų struktūrą, gaunamas tinklapių klasteris, kuriame kiekvienas iš tinklapių „reklamuoja“ ir palaiko kitus (24 pav.).



24 pav. ***Kompanijos „Expo-Vakarai“ tinklapių klasteris (2005-05-10). Pilkos strėlės yra nuorodos.***

http://www.touchgraph.com/TGGoogleBrowser.php?start=www.profis.lt

5. PAIEŠKOS SISTEMŲ VYSTYMOSI PERSPEKTYVOS.

D. Lande, paieškos sistemų specialisto, tvirtinimu, efektyvi paieška, laiku pasiūlanti reikalingas žinias ir pasižyminti savybe išvengti informacinio triukšmo, tampa pagrindiniu faktoriumi didinant konkurencingumą. Be paieškos ir teksto analizės sistemų bet kokie rimti informaciniai projektai rytoj bus pasmerkti [22]. „Paieškos sistemų rinka paskutiniaisiais metais nepaprastai išaugo, - komentuoja kompanijos „JupiterResearch“ analitikas Neitas Eliotas (Nate Elliott). – Dabar ji susiformavusi pakankamai“ [21].

Tradicinės paieškos technologijos, pagristos loginių operatorių panaudojimu, žlugo kartu su web-technologijų bumu. Visos viltys remtis vartotojų intelektu formuluojant užklausas buvo beprasmės. Todėl šiuolaikinės paieškos sistemos pačios vis labiau intelektualizuojasi, turi semantinius įrankius, stengiasi išryškinti vartotojų informacinius poreikius ir remiasi jais paieškoje. Pirmu smuiku paieškos sistemose gros ne DB indeksavimo instrumentai ir loginės paieškos organizacija, bet nauji semantiniai algoritmai. Galima teigti, jog šioje srityje pionieriu galima laikyti kompaniją **Google**, kuri skyrė visas jėgas pertinentiško reitingavimo algoritmų tyrimams. Ir ateitis priklauso tokiems algoritmams.

Paskutiniu metu į paieškos sistemų technologijas vis dažniau įterpiami „Text Mining“ (gilios lingvistinės ir semantinės teksto analizės) elementai:

- Automatiškas dokumentų grupavimas pagal klasifikatorių.
- Automatinis naujų, iš anksto nenusakytų, klasių nustatymas nestruktūrizuotų arba silpnai struktūrizuotų dokumentų pagrindu.
- Dokumentų reitingavimas pagal pertinentiškumą.
- Semantiškai panašių dokumentų nustatymas ir dokumentų paieška etalono pagrindu.
- Automatiška analizė ir prasminis vartotojų užklausų pakeitimas.

IŠVADOS IR TOLIMESNIŲ DARBŲ MATMENYS

1. Išanalizavus pasaulinio tinklo informacijos struktūros dinamiką ir naujus vartotojų poreikius, konkrečiais pavyzdžiais buvo pademonstruota paieškos sistemų svarba šiuolaikiniam verslui ir informacinei visuomenei.

2. Ištyrus šiuo metu populiariausios paieškos sistemos Google bei kitų paieškos sistemų ir katalogų veikimo principus, buvo aprašyti jų funkcionavimo ypatumai ir nustatyti ryšiai tarp jų.

3. Išanalizavus tinklapių reitingo paieškos sistemose sudarymo algoritmus, buvo pasiūlytas apytikslis reitingo skaičiavimo metodas, jo pagrindu suformuluota tinklapių optimizavimo paieškos sistemoms metodika bei pasiūlyti praktinės rekomendacijos tinklapių kūrėjams.

4. Darbe pasiūlytos strategijos ir metodai yra praktiškai įgyvendinti daugiau kaip 60 kompanijos ProfIS vykdytų projektų.

Ateityje ketinama tęsti tyrimus tokiose šios srities kryptyse:

- „Text Mining“ technologijų taikymas paieškos sistemose.
- Automatinis raktinių žodžių generavimas atsižvelgiant į lietuvių kalbos semantiką.
- Lietuviško Interneto (LITneto) dinamikos tendencijų stebėjimas ir analizė.

LITERATŪROS SĄRAŠAS

[1] Acharya A., Cutts M., Dean J. ir kt. *Information retrieval based on historical data.* US Patent N20050071741, CC: 715/500, IC: G06F 017/00, March 31, 2005 [interakyvus]. (žiūrėta 2005 05 16), http://appft1.uspto.gov/netacgi/nph-Parser?Sect1=PTO2&Sect2=HITOFF&p=1&u=%2Fnetahtml%2FPTO%2Fsearch-bool.html&r=1&f=G&l=50&co1=AND&d=PG01&s1=20050071741&OS=20050071741&RS=20050071741

[2] Baldi P., Frasconi P., Smyth P. *Modeling the Internet and the Web: Probabilistic Methods and Algorithms*, Wiley, 2003. 306 p. ISBN: 0-470-84906-1

[3] Banks M. *Search Engine Submission in a Changed Search World*. Iš „Search Engine Optimism“ tinklapio [interakyvus], 2005. (žiūrėta 2005 04 15), http://searchengineoptimism.com/search_engine_submission.html

[4] Barroso L. A., Dean J., Holzle U. *Web Search for a Planet: The Google Cluster Architecture*. Iš “IEEE Computer Society's Web” tinklapio [pdf], 2003. (žiūrėta 2005 04 18), http://www.computer.org/micro/mi2003/m2022.pdf

[5] Beasley C. *Urban Myths - List of urban myths about Google*, Iš „Submit to Google“ tinklapio [interakyvus]. (žiūrėta 2005 05 15), http://www.stevenforsyth.com/google/urban-myths.html

[6] Beckett H. *Web links lead to dead-ends.* Iš “ComputerWeekly.com” portalo [interaktyvus]. 2000. (žiūrėta 2005 04 06), http://www.computerweekly.com/Article20190.htm

[7] Berners-Lee T., Hendler J., Lassila O. *The Semantic Web: A new form of Web content that is meaningful to computers will unleash a revolution of new possibilities*. Iš “Scientific American.com” portalo [interaktyvus]. JAV, 2001. (žiūrėta 2005 04 01), http://www.sciam.com/article.cfm?articleID=00048144-10D2-1C70-84A9809EC588EF21

[8] Brin S., Page L. *The Anatomy of a large scale hypertextual web search engine* [pdf], JAV, CA: Stanford Unversity, 1998. (žiūrėta 2004 04 20), http://www-db.stanford.edu/pub/papers/google.pdf

[9] Bruce Clay, *Internet Business Consultants* [interaktyvus]. JAV, CA: Bruce Clay Inc., 2005. (žiūrėta 2005 01 22), http://www.bruceclay.com/

[10] Buresh S. *Are Your Search Engine Rankings At Risk?*, Iš „Search Engine Guide“ tinklapio, [interakyvus], 2004. (žiūrėta 2005 05 15),

http://www.searchengineguide.com/buresh/2003/0224_sb1.html

[11] Chakrabarti S., Dom B., Gibson D. ir kt. *Hypersearching the Web.* Iš „Scientific American" elektroninio leidinio [interaktyvus]. JAV, NY: Cornell University, 1999. (žiūrėta 2005 04 06), http://www.cs.cornell.edu/home/kleinber/sciam99.html

[12] Craven Ph. *Google's PageRank Explained*. Iš "WebWorkshop" tinklapio [interakyvus], UK, 2004. (žiūrėta 2005 04 16), http://www.webworkshop.net/pagerank.html

[13] Craven Ph. *Inbound links, link exchanges and link acquisition*. Iš "WebWorkshop" tinklapio [interakyvus], UK, 2004. (žiūrėta 2005 04 16), http://www.webworkshop.net/inbound-links.html

[14] Google, *Technology Overview* [interakyvus], 2004. (žiūrėta 2005 04 16), http://www.google.com/corporate/tech.html

[15] Junghoo C., Sourashis R. *Impact of Search Engines on Page Popularity*. Iš UCLA tinklapio [pdf], JAV, LA: University of California, 2004. (žiūrėta 2005 03 09), http://oak.cs.ucla.edu/~cho/papers/cho-bias.pdf

[16] Kaleka H. *Keyword Research for Search Engine Optimization.* Iš „RedAlkemy" kompanijos tinklapio [interakyvus], 2004. (žiūrėta 2005 05 12), http://www.redalkemi.com/search-engine-optimization-seo/keyword-research-article.php

[17] Kenkai WebDesign, *Robots Exclusion Protocol* [interakyvus], 2004. (žiūrėta 2005 05 16), http://www.kenkai.com/robots-txt-exclusion-protocol.htm

[18] Khalil A., Liu Y. *Experiments with PageRank Computation*, Iš "Indiana University" tinklaipio [interakyvus], JAV, 2003. (žiūrėta 2005 02 15), http://informatics.indiana.edu/fil/Class/b659/Projects/S04-g2/main.htm

[19] Kleinberg K., Lawrence S. *The Structure of the Web*. Elektroninis leidinis „Science's Compass" [pdf], American Association for the Advancement of Science, vol. 294, 2001. ISSN 1095-9203. (žiūrėta 2005 04 09), http://www.cs.cornell.edu/home/kleinber/sci01.pdf

[20] LaMonica M. *Google's Secrets of success.* Iš "CNET News.com" tinklapio [interakyvus], 2005. (žiūrėta 2005 04 15), http://news.com.com/Googles+secret+of+success+Dealing+with+failure/2100-1032_3-5596811.html

[21] Lande D. *Кому в сети жить хорошо?* Iš D. Lande tinklapio [interakyvus], 2004. (žiūrėta 2005 05 11), http://dwl.visti.net/art/msbiz/index1.html

[22] Lande D. *Поиск знаний в Internet*, Москва: Диалектика, 2005. 271 p. ISBN 5-8459-0764-0.

[23] Lande D. *Поисковые системы: поле боя – семантика.* Iš „InfoStream“ tinklapio [interaktyvus], 2004. (žiūrėta 2005 02 09), http://infostream.com.ua/publ/ast/

[24] Monash Information Services [interakyvus], 2005. (žiūrėta 2005 04 16), http://www.monash.com

[25] Mudd G. *Press release for April 2004*, Analitinės kompanijos „comScore Networks“ tinklapis [interakyvus]. (žiūrėta 2005 05 15), http://www.comscore.com/press/release.asp?id=464

[26] Netcraft, *April 2005 Web Server Survey*. Analitinės kompanijos “Netcraft” tinklapis [interakyvus], Didžioji Britanija, 2005. (žiūrėta 2005 05 19), http://news.netcraft.com/archives/web_server_survey.html

[27] Nielsen. J. *When Search Engines Become Answer Engines.* Elektroninis žurnalas “Alertbox” [interakyvus], August 16, 2004. ISSN 1548-5552. (žiūrėta 2005 03 06), http://www.useit.com/alertbox/20040816.html

[28] Nielsen NetRatings, *The global standarts for Internet audience measurement and anglysis* [interakyvus], 2005. (žiūrėta 2005 05 06), http://www.nielsen-netratings.com/

[29] Nobles R., Gredy K. L. *Web Site Analysis and Reporting*, USA, Roseville CA: Prima Publishing, 2001. 633 p. ISBN: 0-7615-2842-3.

[30] Olsen S. *Does search engine's Power threaten web's independence?* Iš “CNET News.com“ tinklapio [interakyvus], 2002. (žiūrėta 2005 04 16), http://news.com.com/2009-1023-963618.html

[31] Orange Peel, *The History of Internet* [interakyvus], 2003. (žiūrėta 2005 05 19), http://www.orangepeel.com/newOP/en/internet/shortHistory.php?arpanet

[32] Ostashevskij G. *Оптимизация: осторожно передозировка!* Elektroninis leidinis “WebInform” [interakyvus], 2003. СМИ Эл. № 77-8363. (žiūrėta 2005 05 05), http://www.webinform.ru/opinion/320.html

[33] Pennock D. M., Flake G.W., Lawrence S., Glover E. J. *Characterizing the competition for links on the web*. Elektroninis leidinis “PNAS” [pdf], vol. 99, no. 8, JAV: Stanford University's Highwire Press, 2002-. ISSN 1091-6490. (žiūrėta 2005 04 02), http://www.pnas.org/cgi/reprint/99/8/5207.pdf

[34] SearchEngines.ru, *Энциклопедия поисковых систем* [interaktyvus], 2005. (žiūrėta 2005 04 02), http://www.searchengines.ru/

[35] SearchEngineWatch [interaktyvus], Jupitermedia Corporation, 2005. (žiūrėta 2005 01 17), http://www.searchenginewatch.com

[36] Sobek M. *Google Dance - The Index Update of the Google Search Engine*. Iš „eFactory“ tinklapio [interakyvus], Vokėtija, 2003. (žiūrėta 2005 04 16), http://dance.efactory.de/

[37] Sobek M. *Google's PageRank 0 Penalty*. Iš „eFactory“ tinklapio [interakyvus], Vokėtija, 2002. (žiūrėta 2005 02 16), http://pr.efactory.de/e-pr0.shtml

[38] Spring T. *Three Minutes With Google's Eric Schmidt*. Iš „PCworld“ portalo [interakyvus], 2002. (žiūrėta 2005 03 21), http://www.pcworld.com/news/article/0,aid,81685,00.asp

[39] Technical University of Denmark, *MOLE: Text analysis group* [interaktyvus], 1999. (žiūrėta 2005 04 04), http://isp.imm.dtu.dk/thor/projects/multimedia/textmining/index.html

[40] Tihonov V. *Архитектура метапоисковых систем*. Iš “Indexator.ru” tinklapio [interakyvus], 2004. (žiūrėta 2005 02 12), http://lib.indexator.ru/article/3948

[41] University of Alberta, *Markovo Modelis* [ppt], Canada, 2005. (žiūrėta 2005 03 13), www.cs.ualberta.ca/~lindek/366/slides/MarkovModel.ppt

[42] Wikipedia, *Search engine optimization* [interaktyvus], 2005. (žiūrėta 2005 02 06), http://en.wikipedia.org/wiki/Search_engine_optimization

[43] Wikipedia, *The Google Bomb* [interaktyvus], 2005. (žiūrėta 2005 02 07), http://en.wikipedia.org/wiki/Google_bomb

[44] Zolotov E. *Google сделала это.* Iš “Indexator.ru” tinklapio [interakyvus], 2004. (žiūrėta 2005 02 12), http://lib.indexator.ru/article/3923

Darbe taip pat buvo panaudota informacija iš šių forumų:

SEO Chat: http://forums.seochat.com/
SpiderFood: http://forums.spider-food.net/
SEO Forum: http://www.webworkshop.net/seoforum/index.php
Hight Ranking: http://www.highrankings.com/forum/
Search Guild: http://www.searchguild.com
WebMaster World: http://www.webmasterworld.com/
SEOforum.com: http://www.seoforum.com/
BlueGecko: http://www.bluegeckonetwork.com/forum/
Michigan Web Solutions: http://www.mi-websolutions.com/forum/
SeoGuy: http://www.seo-guy.com/forum/
Cre8asite Forums: http://www.cre8asiteforums.com/
Best Practices Search Engine Forums: http://www.ihelpyouservices.com/forums/
Talk SEO: http://www.searchengineforums.com/
Search Engine Marketing: http://www.top25web.com/bbs/index.php
Seo-Forums.com: http://www.seo-forums.com/
SEO Project: http://www.seoproject.com/forum/
Search Engine Watch: http://searchenginewatch.com/searchday/article.php/3418841
Digital Point: http://forums.digitalpoint.com/
V7Network Forums: http://www.v7n.com/forums/
AKA Marketing: http://www.akamarketing.com/webmaster-forums/

PRIEDAS 1: SEO PROGRAMŲ SĄRAŠAS:

- **SEOMonitor** – kompleksinis SEO programų paketas, susidedantis iš tokių įrankių: tinklapių pozicionavimas, tinklapių indeksacija, tinklapio lankomumo analizė, raktinių žodžių analizė, HTML-analizatorius.
  http://www.semonitor.ru/

- **Google Toolbar 3** – IE papildomų įrankių juosta, vaizduojanti tinklapio PR.
  http://toolbar.google.com

- **Maximized Software FlashStats** – tinklapio lankomumo analizė.
  http://www.maximized.com

- **Surfstats Log Analyzer** – tinklapio žurnalo bylos (log-file) analizė.
  http://www.surfstats.com

- **WebTrends CommerceTrends** – e-verslas, lankomumo analizė, tinklapio pozicionavimas.
  http://www.webtrends.com

- **Search Engine Optimizer** – tinklapio optimizacija.
  http://www.se-optimizer.com

- **Top Dog** – tinklapių pozicionavimas ir registravimas paieškos sistemose.
  http://www.topdog2000.com

- **Web Position Gold** – tinklapių pozicionavimas ir registravimas paieškos sistemose.
  http://www.firstplacesoftware.com

- **Page Promoter** – programinių tinklapių optimizacijos produktų paketas, apimantis visas tinklapių optimizavimo kryptis.
  http://net-promoter.com/ru/pagepromoter/

- **Free Monitor for Google** – tinklapių pozicionavimo stebėjimai pagal raktinių žodžių sąrašą.
  http://www.cleverstat.com/google-monitor-query.htm

PRIEDAS 2: PAIEŠKOS ROBOTŲ SĄRAŠAS

Paieškos sistema	Referer	IP adresai
www.abacho.com	AbachoBOT	srv-ze-robot1.tricus.com
www.abcdatos.com	abcdatos_botlink	217.126.39.167
www.aesop.com	AESOP_com_SpiderMan	209.189.115.49
www.ah-ha.com	ah-ha.com crawler	c7pub-216-250-141-186.center7.com
www.alexa.com	ia_archiver	green.alexa.com, sarah.alexa.com
www.altavista.com	Scooter Mercator Scooter2_Mercator_3-1.0 roach.smo.av.com-1.0 Tv**_Merc_resh_26_1_D-1.0	test-scooter.pa.alta-vista.net, tv**.sv.av.com brillo.pa.alta-vista.net, av-dev4.pa.alta-vista.net scooter.aveurope.co.uk, bigip1-snat.sv.av.com mercator.pa-x.dec.com, scooter.pa.alta-vista.net scooter.sv.av.com, avfwclient.sv.av.com
www.altavista.co.uk	AltaVista-Intranet	host-119.altavista.se
www.alltheweb.com	FAST-WebCrawler	209.67.247.154
www.acoon.de	Acoon Robot	194.231.42.178
www.antisearch.net	antibot	62.210.155.50
www.axmo.com	AxmoRobot	194.248.208.82
www.canseek.ca	CanSeek	216.168.111.111
www.christcrawler.com	ChristCRAWLER	207.191.111.231
www.clush.com	Clushbot	209.249.80.242
www.crawler.de	Crawler	crawlit.crawler.de
www.daadle.com	DaAdLe.com ROBOT/	216.12.213.32
www.daum.net	RaBot	210.183.28.46, 211.50.57.6, 202.30.94.34
www.en.deepindex.com	DeepIndex	deepindex.net1.nerim.net
www.delfi.lt	FAST Enterprise Crawler/6 (crawl at delfi dot lt)	www.delfi.lt
www.ditto.com	DittoSpyder	65.169.94.188
www.earthcom.info	EARTHCOM.info	194.108.39.74
www.entireweb.com	Speedy Spider	62.13.25.209
www.excite.com	ArchitextSpider	********.excite.com
www.eurip.com	EuripBot	81.169.172.30
www.euroseek.net	Arachnoidea	212.209.54.134
www.ezresults.com	EZResult	216.28.23.59
www.fybersearch.com	FyberSearch	69.49.241.9
www.galaxy.com	GalaxyBot	63.121.41.175
www.geckobot.com	geckobot	***.rdc1.az.coxatwork.com
www.geona.com	GeonaBot	69.59.142.17
www.getrax.com	getRAX	81.169.156.246
www.google.com	Googlebot/2.1	c**.googlebot.com
www.girafa.com	Aranha	Aranha.girafa.com
www.inktomi.com	Slurp.so/1.0, Slurp/2.0j Slurp/2.0-KiteHourly Slurp/2.0-OwlWeekly	q2004.inktomisearch.com, j5006.inktomisearch.com 202.212.5.34, goo313.goo.ne.jp y400.inktomi.com 209.185.143.198, j6000.inktomi.com
http://hoppa.com/	Toutatis 2.5-2	tisnix.xs4all.nl
www.hubat.com	Hubater	209.114.176.250
www.incywincy.com	IncyWincy	64.81.243.66
www.infoseek.com	UltraSeek InfoSeek Sidewinder	cde2c923.infoseek.com, cde2c91f.infoseek.com, cca26215.infoseek.com

Paieškos sistema	Referer	IP adresai
www.intags.de	Mole2/1.0	217.160.75.10
www.ip3000.com	ip3000.com-crawler	www.ip3000.com
www.knowledge.com	Knowledge.com/	213.170.2.69
www.kuloko.com	kuloko-bot/0.2	66.90.81.41
www.lexis-nexis.com	LNSpiderguy	firewall5.lexis-nexis.com
www.linknz.co.nz	Linknzbot	202.191.32.67
www.look.com	lookbot	magma.com
www.looksmart.com	MantraAgent	fjupiter.looksmart.com
www.loopimprovements.com	NetResearchServer	leg-64-133-109-250-STK.sprinthome.com
www.lycos.com	Lycos_Spider_(T-Rex)	bos-spider*.bos.lycos.com, 216.35.194.188
www.joocer.com	JoocerBot	80.46.38.169
www.mirago.co.uk	HenryTheMiragoRobot	194.202.39.46
http://search.msn.com/	MSNBOT/0.1	131.107.163.47
www.northernlight.com	Gulliver	marvin.northernlight.com, taz.northernlight.com
www.objectssearch.com	ObjectsSearch/0.01	68.88.244.177
www.picosearch.com	PicoSearch/	pipe.picosearch.com
www.portaljuice.com	PJspider	timber.nextopia.com
www.powerinter.net	DIIbot	node-d8e93393.powerinter.net
www.maxbot.com	Spider/maxbot.com	search.wport.com
www.naver.com	dloader(NaverRobot)	211.218.151.209
www.openfind.com	Openbot	abovenet4.openfind.com
www.picsearch.org	psbot	217.75.104.26
www.pinpoint.com	CrawlerBoy Pinpoint.com	nitrogen.pinpoint.com
www.petersnews.com	user*.ip3000.com	news*.petersnews.com
www.qweery.nl	QweeryBot	84.82.133.41
www.vestris.com/alkaline	AlkalineBOT	host130.uv-ray.com
www.search-10.com	Search-10	82.41.144.99
www.searchhippo.com	Fluffy the spider	208.148.122.27
www.scrubtheweb.com	Scrubby/	208.145.190.254
www.singingfish.com	asterias	grouper.singingfish.com
www.searchspider.com	Searchspider/	24.90.243.203
www.sightquest.com	SightQuestBot/	64.49.245.212
www.spidermonkey.ca	Spider_Monkey/	66.163.18.197
www.surfnomore.com	Surfnomore Spider v1.1	165.90.194.245
www.supersnooper.com	Robot@SuperSnooper.Com	207.8.212.162
www.teoma.com	teoma_agent1	63.236.92.148
http://mapper.teradex.com	Teradex_Mapper	65.110.6.26
www.travel-finder.com	ESISmartSpider	202.46.33.15
www.turnitin.com	TurnitinBot/2.0	www.turnitin.com
www.tutorgig.com	Tutorial Crawler	216.40.225.75
www.updated.com	updated/0.1beta	38.119.96.107
www.webwombat.com.au	www.WebWombat.com.au	202.139.99.131
www.webseek.de	marvin/infoseek	arthur4.sda.t-online.de
www.webtop.com	MuscatFerret	ferret**.webtop.com
www.whizbanglabs.com	WhizBang! Lab	216.250.143.108
www.yandex.ru	Yandex	ya.yandex.ru
www.yahoo.com	Slurp	www.yahoo.com