KAUNO TECHNOLOGIJOS UNIVERSITETAS
INFORMATIKOS FAKULTETAS
PROGRAMŲ INŽINERIJOS KATEDRA

Internetinio paieškos proceso modelio kūrimas ir tyrimas

Magistro darbas

Užsakovas: prof. Vacius Jusas

Vadovas: prof. Vacius Jusas

Autorius: Augustinas Gustas

Kaunas, 2011

KAUNO TECHNOLOGIJOS UNIVERSITETAS
INFORMATIKOS FAKULTETAS
PROGRAMŲ INŽINERIJOS KATEDRA

Internetinio paieškos proceso modelio kūrimas ir tyrimas

Magistro darbas

Recenzentas:		**Vadovas:**	
2011-05-30		**2011-05-30**	**Dr. Vacius Jusas**
		Atliko:	**IFM-5/2 gr. stud.**
		2011-05-30	**Augustinas Gustas**

Kaunas, 2011

Turinys

Implementation and Research of Web Search Process Model

Summary

The main objective of this project was to develop the web search process model (web crawler), which allows to scan and analyse specified websites, download their content, extract links and documents. During the execution of this project the analysis of design and technology solutions was performed. The architecture of the developed system is based mainly on the principles of client server and MVC (Model–view–controller) software designs.

In the research and experimentation part of this document, there was performed analysis of various website quality metrics. These metrics allows web crawler to evaluate the quality of specified websites and then organize deeper analysis of these websites in an efficient way.

Paveikslų sąrašas

Lentelių turinys

1. Įvadas

Temos aktualumas

Informacinės technologijos šiais laikais yra tapę neatsiejama gyvenimo dalimi. Įvairios technologijos bei programinės sistemos palengvina mūsų įprastinį gyvenimą bei padidina darbo kokybę ir našumą.

Gyvename informaciniame amžiuje, todėl kasdien įvairių šaltinių pagalba mus pasiekia didžiuliai informacijos kiekiai. Vienas pagrindinių ir didžiausių informacijos šaltinių yra internetas. Internete pateikiamos informacijos kiekiai nuolat auga, todėl vis aktualesne problema tampa svarbios bei reikiamos informacijos paieška bei išrinkimas.

Informacijos analizei bei rinkimui internete yra naudojamos sistemos, vadinamos paieškos robotais. Šios sistemos yra autonominės: jos savarankiškai klaidžioja po internetą, ieško naujų svetainių bei nuorodų, analizuoja bei kaupia svetainėse pateikiamą turinį. Ši sukaupta ir susisteminta informacija vėliau yra panaudojama įvairiais tikslais: viena pagrindinių panaudojimo sričių – paieškos sistemos. Paieškos sistemos pagal įvestus raktinius žodžius vartotojui pateikia paieškos robotų sukauptą aktualią informaciją.

Atliekant paiešką internete svarbia problema tampa informacijos šaltinių – internetinių svetainių – svarbos nustatymas. Svetainių nuorodų svarbos koeficientų nustatymas leidžia paieškos procesui pirmiau nagrinėti aukštesnės svarbos svetaines, taip pasiekiant didesnį paieškos proceso efektyvumą bei geresnius rezultatus.

Darbo tikslas

Pagrindinis magistrinio darbo tikslas sukurti internetinės paieškos procesą (paieškos robotą). Ši sistema gali būti naudojama pasirinktų internetinių svetainių turinio analizei, vidinių / išorinių nuorodų bei dokumentų paieškai, pageidaujamos informacijos išrinkimui ir išsaugojimui.

Magistrinio darbo metu taip pat siekiama detaliai išnagrinėti ir susipažinti su internetinės paieškos procesų kūrimo principais bei metodais, dažniausiai pasitaikančiomis problemomis ir jų sprendimo būdais.

Dokumento paskirtis

Šiame baigiamojo magistrinio darbo dokumente aprašomi internetinės paieškos proceso kūrimo principai, naudojamų paieškos sistemų tipai, egzistuojantys paieškos robotų architektūros sprendimai. Dokumentas taip pat supažindina skaitytoją su sukauptų duomenų analize, svetainių

svarbos nustatymo procesu. Taip pat aptariamos dažniausiai praktikoje pasitaikančios paieškos proceso problemos ir rekomenduojami jų sprendimo būdai.

Dokumento struktūra

Šiame dokumente informacija pateikiama struktūrizuotai, išskaidžius informaciją į atskirus skyrius. Kiekviename skyriuje nagrinėjama viena konkreti tema. Žemiau pateikiamas pagrindinių šio dokumento skyrių sąrašas bei trumpi jų aprašymai.

- Analitinė dalis. Šiame skyriuje pateikiama literatūros analizė, apibrėžiama internetinės paieškos proceso paskirtis bei panaudojimo galimybės. Nagrinėjami paieškos proceso architektūros modeliai. Taip pat aprašomos dažniausiai pasitaikančios problemos ir jų sprendimo būdai.
- Projektinė dalis. Šiame skyriuje pateikiama informacija apie realizuotą internetinės paieškos procesą. Čia aprašomi sistemos funkciniai bei nefunkciniai reikalavimai, pateikiamas sistemos panaudojimo atvejų sąrašas. Šioje dalyje paaiškinami pasirinkti sistemos architektūros sprendimai, apibrėžiama duomenų struktūra, pagrindiniai sistemos komponentai.
- Tyrimo dalis. Šiame skyriuje aprašomas magistrinio darbo tyrimas, jo reikalingumas. Tyrimo metu buvo atlikta svetainių svarbos nustatymo metrikų analizė.
- Eksperimentinė dalis. Šioje darbo dalyje aprašomi magistrinio darbo metu atlikti eksperimentai bei jų rezultatai.
- Išvados. Šiame dokumento skyriuje pateikiamos esminės magistrinio darbo metu gautos išvados bei rezultatai.

Kiekvieno skyriaus pabaigoje pateikiamos pagrindinės to skyriaus išvados.

2. Analitinė dalis

2.1. Literatūros apžvalga

Šiame skyriuje pateikiama magistrinio darbo temos – internetinio paieškos proceso modelio kūrimas ir tyrimas – literatūros apžvalga. Detalų analizės metu nagrinėtų literatūros šaltinių sąrašą galite rasti 7 skyrelyje „Literatūra“.

2.1.1. Internetinis paieškos procesas

Internetinis paieškos procesas (paieškos voras) – tai savarankiška sistema, klaidžiojanti internete, atsitiktine arba nustatyta eiliškumo tvarka nagrinėjanti internetines svetaines bei jų turinį. Svetainių eiliškumo nustatymas remiasi pasirinktomis svarbos įvertinimo metrikomis.

Pirmasis internetinis paieškos procesas buvo sukurtas 1993 metais Masačiuseto Technologijų universiteto studento Matthew Gray [1]. Ši sistema buvo naudojama naujų svetainių paieškai bei interneto augimo matavimui. Sistema veikė nuo 1993 iki 1996 metų [2].

Šiuo metu pasaulyje egzistuoja daugybė komercinių, atvirojo kodo, privačių asmenų bei organizacijų sukurtų paieškos procesų. Plačiau apie egzistuojančius paieškos procesus galite skaityti skyriuje 2.1.4. „Egzistuojantys sprendimai“.

2.1.2. Paieškos proceso veikimo principai

Internetinio paieškos proceso veikimo principai iš esmės priklauso nuo jo paskirties, tačiau galima išskirti šiuos bendrinius žingsnius, būdingus daugumos tokio tipo sistemų veikimui [3]:

- Paieškos procesui pateikiamas konkretus pradinis svetainių adresų rinkinys. Šis nuorodų rinkinys yra sudaromas sistemos vartotojo.
- Paieškos procesas rekursiškai lanko visas svetainės nuorodas patikrindamas jų egzistavimą bei sėkmės atveju išsaugodamas lankomo svetainės puslapio turinį. Svetainių nuorodų lankymo tvarka gali būti atsitiktinė arba paremta tam tikromis nustatytomis nuorodų svarbos koeficientų reikšmėmis.
- Atliekama išsaugotų svetainės puslapių turinio analizė siekiant išrinkti visas nuorodas į vidinius puslapius bei išorines svetaines. Taip pat gali būti išrenkama ir išsaugojama kita pasirinkta puslapio turinio informacija. Tada atliekami veiksmai aprašyti ankstesniame žingsnyje.

Žemiau pateikiamas internetinio paieškos proceso veikimo pseudo kodo pavyzdys [4]:

```
Ask user to specify the starting URL on web and file type that crawler should crawl.
Add the URL to the empty list of URLs to search.

While not empty ( url list )
{
        Take the first URL in from the list of URLs
        Mark this URL as already searched URL.

        If the URL protocol is not HTTP then
            break;
            go back to while

        If robots.txt file exist on site then
            If file includes “Disallow” statement then
           break;
           go back to while

        Open the URL
        If the opened URL is not HTML file then
            Break;
            Go back to while

        Iterate the HTML file

        While the html text contains another link {
            If robots.txt file exist on URL/site then
                If file includes “Disallow” statement then
                break;
               go back to while

            If the opened URL is HTML file then
                If the URL isn't marked as searched then
                    Mark this URL as already searched URL.

           Else if type of file is user requested
               Add to list of files found.
        }
}
```

2.1.3. Paieškos proceso panaudojimo galimybės

Internetinio paieškos proceso taikymo galimybės yra labai plačios. Ši sistema gali būti naudojama siekiant automatizuoti bei palengvinti tam tikrus specifinius organizacijos darbus, kaip pvz: svetainių kodo validavimas, skirtingų šaltinių naujienų srauto sekimas. Šiuo metu taip pat intensyvėja paieškos sistemų naudojimas kenkėjiškais tikslais: asmens privačių duomenų rinkimas,

turinio vagystės be autoriaus sutikimo. Žemiau pateikiamas populiariausių paieškos roboto panaudojimo galimybių sąrašas [4].

- Specializuotų katalogų sudarymas. Paieškos procesas gali būti naudojamas siekiant sudaryti tam tikros specialios paskirties svetainių / nuorodų katalogą.
- Autorinės apsaugos palaikymas. Internetinė paieška gali būti vykdoma siekiant nustatyti nelegalaus turinio kopijavimo, autorinių teisių pažeidimo atvejus.
- Svetainių naujienų sekimas. Paieškos procesas gali sekti pasirinktų svetainių turinio bei struktūros pasikeitimus.
- Svetainės kodo validavimas. Paieškos procesas gali būti naudojamas siekiant rasti neveikiančias svetainės nuorodas, validuoti svetainės HTML kodą.
- Svetainių kopijavimas. Suteikiama galimybė pasidaryti identišką pasirinktos svetainės kopiją. Išsaugoma visa svetainės turinio bei nuorodų struktūros informacija.
- Specifinės informacijos paieška. Paieškos procesas gali būti naudojamas specifinės informacijos išrinkimui, pavyzdžiui: klientų informacijos, vartotojų elektroninio pašto adresų rinkimas ir kaupimas marketingo tikslais.

2.1.4. Egzistuojantys sprendimai

Pasaulyje šiuo metu egzistuoja daugybė įvairios paskirties komercinių, atvirojo kodo, privačių asmenų bei organizacijų sukurtų ir naudojamų paieškos procesų (paieškos vorų). Remiantis [5] šaltiniu, šiuo metu užregistruota 5275 tokio tipo sistemos.

Visi paieškos procesai vienas nuo kito skiriasi savo architektūriniais sprendimais, taikoma paieškos strategija, galimybėmis, veiklos efektyvumu bei informacijos rinkimo paskirtimi. Žemiau pateikiami didžiausių pasaulyje paieškos sistemų „Google“, „Yahoo“, „Microsoft Bing“ paieškos vorų aprašymai. Taip pat aprašoma ir keletas pasirinktų smulkesnių tokio tipo sistemų.

Googlebot

“Googlebot” – tai internetinis paieškos procesas naudojamas “Google” paieškos sistemos. Naujus puslapius Google paieškos voras randa naršydamas po jau atrastas svetaines arba po naujas svetaines, pateiktas vartotojų. Googlebot – tai labai spartus ir galingas paieškos procesas, kadangi jis veikia dideliame internetinių serverių tinkle [6]. Siekdamas nustatyti svetainių nuorodų svarbą bei apibrėžti analizės eiliškumą paieškos voras įvertina maždaug 200 įvairių metrikų. Nors šis paieškos procesas turi daug galimybių (išrenka svetainių meta žymas, detaliai išanalizuoja puslapio

turinį, išskiria pastraipų antraštes, nustato paveikslėlių alt atributo reikšmes), tačiau jis šiuo metu neturi galimybės pilnai išanalizuoti kai kurių specifinių tipų svetainių. [7].

Bingbot

"Bingbot" tai internetinis paieškos robotas, sukurtas "Microsoft" korporacijos. Šis paieškos procesas 2010 metų rudenį visiškai pakeitė anksčiau naudotą msnbot paieškos sistemą [8]. Paieškos procesas „Bingbot" kaip ir „Googlebot" neturi galimybės išanalizuoti flash svetainių turinio. Taip pat nerekomenduojama naudoti turinio, pateikiamo vidiniuose HTML iframe objektuose. Vertindamas svetainių svarbą paieškos robotas atsižvelgia į domeno galiojimo laiką, svetainės puslapių užkrovimo laiką, nuorodų struktūros gylį, neveikiančių nuorodų skaičių bei kitas metrikas.

Yahoo! Slurp

"Yahoo! Slurp" [9] tai internetinis paieškos procesas, naudojamas "Yahoo" paieškos sistemos. Šis paieškos robotas kaip ir anksčiau minėtas msnbot naudoja Inktomi paieškos algoritmą. Naršydamas po puslapius šis paieškos procesas atsisiunčia pilną puslapio turinį, kas nėra dažnai naudojama kitose paieškos sistemose. Paminėtina, kad Yahoo paieškos procesas geba efektyviai dirbti ir suindeksuoti dinaminius svetainių puslapius.

Žemiau pateikiami keleto pasirinktų mažesnių paieškos robotų aprašymai bei galimybės, pateiktos [10] literatūros šaltinyje.

WebSphinx – tai paieškos procesas, sukurtas 1998 metais, Milerio ir Barato. Ši sistema sukurta Java programavimo kalbos pagrindu. Sistemoje vienu metu veikia daug lygiagrečių procesų, kas leidžia paspartinti nuorodų analizės bei turinio atsisiuntimo procesą. WebSphinx turi patogią vartotojo sąsają bei savyje turi integruotą nedidelę paieškos sistemą.

Methabot – tai paieškos procesas, realizuotas C kalbos pagrindu, turintis komandinės eilutės sąsają. Šis paieškos robotas yra optimizuotas greitam veikimui, turi modulinę struktūrą, suteikia galimybę reguliuoti didelį skaičių įvairių paieškos proceso parametrų.

Mercator – tai paieškos robotas, sukurtas 1999 metais Heydon ir Nayork. Tai paskirstytos, modulinės architektūros paieškos procesas, parašytas Java programavimo kalba. Sistemos moduliai pagal paskirtį suskirstyti į dvi grupes: turinio atsisiuntimo ir turinio apdorojimo posistemes.

Sistemos pagrindiniai turinio apdorojimo moduliai leidžia tiktai išrinkti nuorodas, papildomi apdorojimo moduliai papildo sistemos funkcionalumą, pavyzdžiui: suteikia galimybę suindeksuoti tekstinę puslapio informaciją bei surinkti puslapio užkrovimo statistiką.

Ubicrawler – paieškos robotas, sukurtas 2002 metais [11]. Sistema parašyta naudojantis Java programavimo kalba, naudojama paskirstyta architektūra, nėra bendro centrinio proceso, reguliuojančio žemesnio lygio paieškos procesus. Šis paieškos robotas suprojektuotas taip, kad būtų pasiekiamas didelis pakeitimų lankstumas bei paieškos proceso patikimumas.

2.1.5. Efektyvaus paieškos proceso savybės

Prieš kuriant internetinį paieškos procesą (paieškos vorą) svarbu apibrėžti esmines tokio tipo sistemų savybes, užtikrinančias sklandų bei efektyvų sistemos darbą. Žemiau pateikiamas sąrašas aspektų bei sistemos kriterijų, į kuriuos rekomenduojama atkreipti dėmesį [12] literatūros šaltinyje.

Lankstumas

Internetinis paieškos procesas turėtų būti lankstus: naujų paieškos scenarijų diegimas neturėtų reikalauti didelių sistemos architektūros pakeitimų. Sistemos lankstumą iš dalies užtikrina modulinė architektūra.

Taupus serverio resursų naudojimas

Egzistuojant daugybei informacijos, pateikiamos internetinėse svetainėse (šiuo metu retai kuri svetainė apsiriboja tik keletu vidinių puslapių), vis aktualesne problema tampa serverio resursų tausojimas. Lankydamas svetainių nuorodas internetinis paieškos procesas turėtų automatiškai nustatyti ribines duomenų kaupimo reikšmes. Tai gali užtikrinti, kad vienos ar kelių svetainių informacija neužims daugiau vietos serveryje nei leidžiama.

Našumas

Internetinis paieškos robotas turi gebėti per sekundę apdoroti ir parsisiųsti didelį kiekį skirtingų internetinių puslapių. Sudėtingoms paieškos sistemoms šis rodiklis siekia keletą tūkstančių puslapių per sekundę.

Patikimumas

Internetinis paieškos procesas – tai autonominė sistema, dirbanti ilgą laiką, todėl svarbu užtikrinti tinkamą sistemos saugumą bei patikimumą. Sistema turi gebėti tinkamai reaguoti į įvairius serverio atsakus užklausoms: jeigu serverio atsakymas yra netinkamas, paieškos robotas tai turėtų užfiksuoti

ir atitinkamai reaguoti. Taip pat svarbu, kad sutrikus sistemos (internetinio voro) veiklai, nebūtų prarandami duomenys sukaupti duomenų saugykloje.

„Etiškumas“ lankomų serverių atžvilgiu

Naršydamas po internetines svetaines, paieškos procesas turi atsižvelgti į tų svetainių draudimus, taikomus paieškos robotams. Nustatymai paprastai pateikiami „robots.txt“ faile [13] arba robotų meta žymose HTML kode. Šie nustatymai gali drausti lankytis tam tikrose internetinės svetainės dalyse arba kataloguose.

Nuolat atnaujindamas informaciją internetinis paieškos robotas gali apkrauti svetaines, kurių puslapius jis atsisiunčia. „Etiška“ sistema turi atsižvelgti į tai, kaip dažnai puslapių turinys pasikeičia ir priklausomai nuo to nustatyti konkrečiam puslapiui informacijos atnaujinimo intervalą.

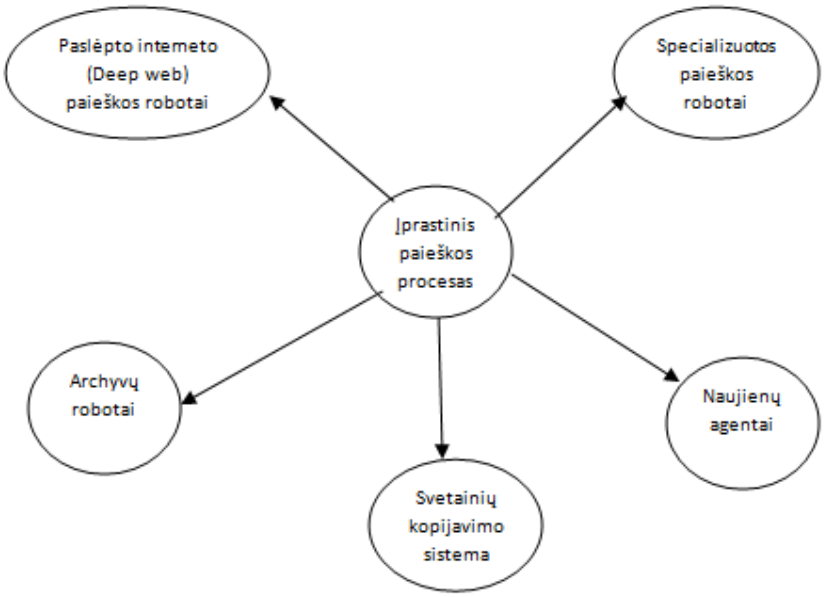
Atlikdamas puslapių atsisiuntimą, internetinis paieškos procesas privalo serveriui pateikti informaciją apie save. Ši informacija apima paieškos roboto kūrėjų kontaktinius duomenis bei paieškos tikslus ir paskirtį.

Patogus valdymas

Svarbu aprūpinti vartotoją galimybe valdyti internetinio paieškos roboto funkcionalumą. Vartotojas turėtų turėti galimybę išjungti paieškos robotą, stebėti pagrindinius veiklos parametrus: duomenų saugyklos užimtumą, puslapių parsisiuntimo greitį ir t.t.

2.1.6. Paieškos procesų (robotų) tipai

Šiuo metu egzistuoja įvairios paskirties internetiniai paieškos procesai (paieškos robotai). Šių sistemų taikymo sritis yra labai įvairi. Žemiau esančiame paveikslėlyje pateikiama diagrama, kurioje atvaizduojami populiariausi paieškos procesų tipai.



2.1 pav. Internetinės paieškos robotų tipai

Įprastiniai paieškos procesai

Tai įprastiniai paieškos robotai, atliekantys naujų svetainių bei nuorodų paiešką. Esant poreikiui suteikiama galimybė išsaugoti nagrinėjamų svetainės puslapių turinį.

Naujienų agentai

Naujienų agentai – tai tokie paieškos robotai, kurių paskirtis sekti svetainių struktūros bei turinio pasikeitimus. Šios sistemos tam tikru laiko intervalu aplanko pasirinktas svetainių nuorodas, palygina esamą puslapių turinį su nuosavoje duomenų bazėje išsaugota informacija. Dažniausiai tokio tipo sistemos naudojamos RSS srauto generavimui.

Svetainių pakartotinio lankymo intervalas yra nustatomas automatiškai (įvertinant įvairius turinio pasikeitimo kriterijus bei atsižvelgiant į svetainės meta žymų informaciją) arba rankiniu būdu (įvedus vartotojui).

Specializuotos paieškos robotai

Specializuoti paieškos robotai yra sukurti rinkti tam tikrą konkrečią informaciją. Tai gali būti straipsniai pasirinkta kalba, paveikslėliai, muzikos failai arba kita specifinė medžiaga. Paieškos procesas paprastai yra atliekamas pagal vartotojo įvestus raktažodžius bei pasirinktus filtravimo kriterijus [14].

Paslėpto interneto (Hidden web) paieškos robotai

Paslėptas internetas – tai tokia interneto sritis, kurios paprastai nesuindeksuoja įprasti internetinės paieškos procesai. Paslėpto interneto (Hidden web) paieškos robotai specializuojasi tokios informacijos išgavimo srityje. Paslėptą informacija gauti yra naudojamas keletas metodų: specialių užklausų pateikimas ir (arba) svetainėje pateiktų formų automatinis užpildymas [15].

Archyvų sudarymo robotai

Archyvų sudarymo robotai specializuojasi tam tikros paskirties informacijos katalogų / nuorodų indeksų sudarymo srityje. Šie robotai kaupia svetainių nuorodas, tenkinančias tam tikrus filtravimo kriterijus. Iš sukauptų duomenų sudaromas katalogas, pavyzdžiui: mokslinių publikacijų archyvas.

Svetainių kopijavimo sistemos

Tai tokie paieškos procesai, kurių tikslas nukopijuoti visą svetainės turinį su vidiniais puslapiais, nuotraukomis, kitomis laikmenomis. Taip pat svetainių kopijavimo sistemos gali palaikyti tam tikrų svetainių rezervines „veidrodines“ kopijas.

2.1.7. Paieškos proceso organizavimo metodikos

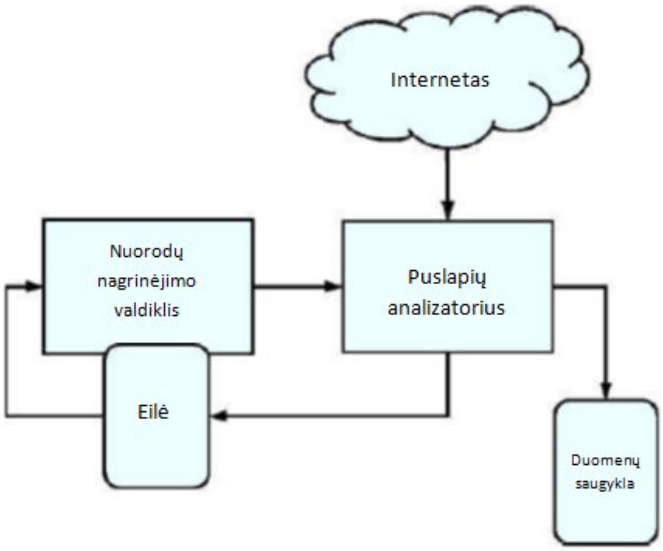
Internetinės paieškos procesai nuorodų nagrinėjimo eiliškumui valdyti taiko įvairaus pobūdžio metodikas. Galima išskirti šias pagrindines bei dažniausiai pasitaikančias paieškos proceso organizavimo metodikas bei strategijas [16]:

- Fiksuota tvarka. Svetainių nuorodos yra lankomos etapais, tačiau kaskart vis ta pačia tvarka. Vieno analizės etapo metu aplankomos visos (arba iki tam tikro gylio esančios) svetainės nuorodos.
- Atsitiktinė tvarka. Nuorodos nagrinėjamos beveik identiškai kaip ir fiksuotos tvarkos paieškos proceso atveju, tačiau nuorodų lankymo tvarka nustatoma atsitiktinai.
- Suteiktos pirmenybės tvarka. Paieškos procesas pirmiau aplanko tas nuorodas, kurioms suteiktas prioritetas. Ši paieškos organizavimo metodika yra labai panaši į svarbos koeficientų metodą (žiūrėti toliau), tačiau šiuo atveju svetainių prioritetus ir svarbą nustato sistemos vartotojai.
- Paieškos platyn metodas. Svetainių nuorodų analizė vykdoma pagal paieškos platyn metodiką. Šiuo atveju paieškos procesas vykdomas paeiliui visoms nuorodoms pagal nuorodos struktūrinį arba suradimo gylį. Užbaigus vieno lygio nuorodų lankymą, pereinama į kitą lygį.

- Paieškos gilyn metodas. Paieškos proceso organizavimas vykdomas remiantis paieškos gilyn metodika. Nuorodų nagrinėjimas vykdomas taip: analizė vykdoma gilyn, kaskart aplankant naują surastą didesnio gylio nuorodą. Paieškos gilyn metodas gali būti reguliuojamas leidžiant sistemai nagrinėti nuorodas tik iki tam tikro gylio.
- Svarbos koeficientų metodas. Kiekvienai svetainei bei nuorodai sistema automatiškai suskaičiuoja jos svarbos koeficientą. Paieškos procesas vykdomas pažingsniui pradedant nuo aukščiausią svarbos reikšmę turinčių nuorodų.

2.1.8. Egzistuojantys architektūros sprendimai

Egzistuoja įvairūs paieškos robotų architektūros sprendimai, tačiau pagrindiniai aspektai yra panašūs visoms sistemoms. Žemiau pateikiamas apibendrintas internetinės paieškos proceso architektūros modelis [17], kuriame galima matyti esminius tokio tipo sistemų komponentus.



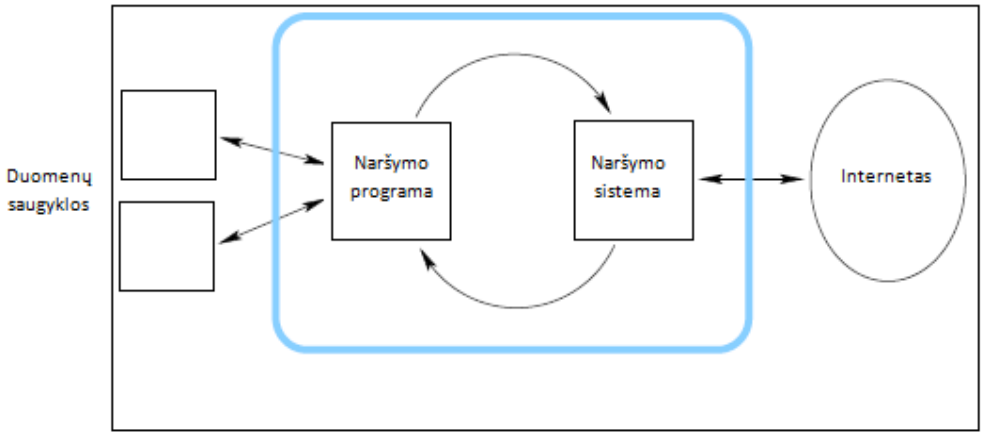
2.2 pav. Bendriniai paieškos proceso architektūros komponentai

Standartinio paieškos proceso architektūros modelis yra sudarytas iš šių komponentų:

- Internetas. Tai pasaulinis interneto tinklas iš kurio paieškos procesas pasiima reikiamą informaciją.
- Puslapių analizatorius. Šis paieškos proceso komponentas yra pagrindinė sistemos dalis, kuri sąveikauja tiek su interneto serveriais, tiek ir su kitais sistemos komponentais, pavyzdžiui: duomenų saugykla, nuorodų nagrinėjimo eiliškumo valdikliu.

- Eilė ir nuorodų nagrinėjimo valdiklis apibrėžia, kurių svetainių / nuorodų analizei suteikti pirmenybę.
- Duomenų saugyklą. Sistemos komponentas, atsakingas už visos paieškos proceso darbui reikalingos informacijos ilgalaikį saugojimą.

Žemiau pateikiamas paieškos sistemos architektūros modelis aprašytas [18] literatūros šaltinyje. Čia sistemos moduliai suskaidomi į dvi pagrindines posistemes: naršymo programos posistemę bei naršymo sistemos posistemę. Abi šios dalys tiesiogiai sąveikauja su pasauliniu interneto tinklu bei paieškos proceso duomenų saugyklomis. Žemiau pateikiama tokios architektūros schema su detalesniais posistemių aprašymais.



2.3 pav. Internetinio paieškos proceso architektūros modelis

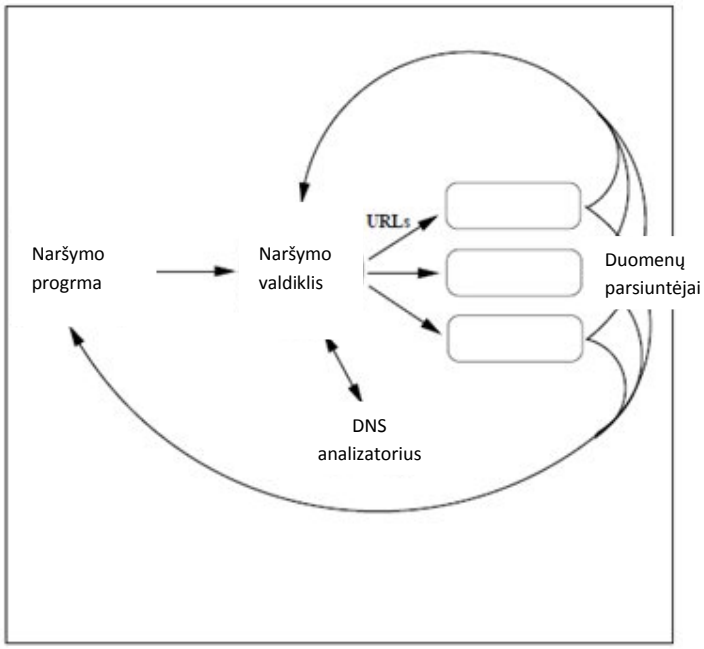
- **Naršymo programa**

Naršymo programa atsakinga už gautų iš naršymo sistemos duomenų analizę. Analizės metu iš puslapio duomenų išskiriami URL adresai, tikrinama ar jie jau aplankyti ar dar ne, galiausiai pageidaujami URL adresai yra perduodami naršymo sistemai. Naršymo programa taip pat atsakinga už paieškos strategijos taikymą bei duomenų išsaugojimą ilgalaikėse paieškos proceso duomenų saugyklose.

- **Naršymo sistema**

Naršymo sistema tiesiogiai bendrauja su pasauliniu interneto tinklu. Ji yra atsakinga už reikiamų puslapių (URL adresus pateikia naršymo programa) atsisiuntimą iš internetinių svetainių. Naršymo sistema taip pat atsakinga už lankomos svetainės „robots.txt“ faile esančių nurodymų laikymąsi bei DNS nagrinėjimą.

Žemiau esančiame paveikslėlyje galima matyti detalizuotą anksčiau pateikto architektūros modelio naršymo sistemos komponento struktūrą. Po paveikslėliu pateikiami pagrindinių šio komponento sudedamųjų dalių aprašymai.



2.4 pav. Detalizuotas paieškos proceso architektūros modelis

- **Naršymo valdiklis**

Valdiklis yra pagrindinė naršymo sistemos dalis. Valdiklis yra atsakingas už duomenų parsiuntėjų bei DNS analizatorių funkcionavimo koordinavimą. Iš naršymo sistemos gavęs pageidaujamus URL adresus, valdiklis nustato kada ir koks URL adresas bus aptarnaujamas duomenų parsiuntėjų bei DNS analizatorių, taip, kad sistemos funkcionavimo efektyvumas būtų didžiausias. Valdiklis taip pat analizuoja „robots.txt“ esančius draudimus paieškos robotams.

- **DNS analizatorius**

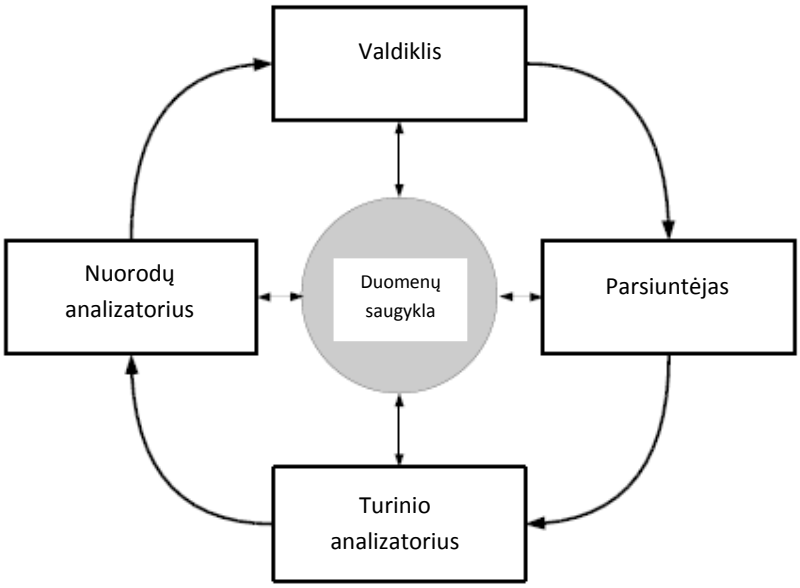
DNS analizatoriaus paskirtis išgauti nagrinėjamų adresų informaciją.

- **Duomenų parsiuntėjas**

Duomenų parsiuntėjas atsako už nurodytų http užklausų pateikimą internetiniam serveriui, serverio atsako registravimą, puslapių informacijos parsisiuntimą. Vėliau ši informacija yra išsaugoma ilgalaikėse duomenų saugyklose.

Žemiau pateikiamas dar vienas galimas internetinio paieškos proceso architektūros modelis pateiktas literatūros šaltinyje [19]. Šis modelis sudarytas iš penkių komponentų. Žemiau pateikiamas komponentų aprašymų sąrašas.

- Duomenų saugykla. Saugo informaciją, reikalingą sistemos darbui.
- Valdiklis. Šis komponentas įvertina bei nustato svetainių nuorodų svarbos koeficientų reikšmes. Valdiklis sugeneruoja sąrašą nuorodų, kurios turi būti aplankytos sistemos komponento „Parsiuntėjas“.
- Parsiuntėjas. Tai sistemos komponentas, atsakingas svetainių nuorodų aplankymą bei jų turinio parsiuntimą ilgalaikiam saugojimui.
- Turinio analizatorius. Sistemos parsiųsta puslapių informacija yra nagrinėjama ir skaidoma turinio analizatoriaus. Čia yra išrenkamos svetainių išorinės bei vidinės nuorodos. Po išrinkimo visos nuorodos yra išsaugomos paieškos proceso duomenų bazėje.
- Nuorodų analizatorius. Čia yra atliekama turinio analizatoriaus išrinktų bei išsaugotų nuorodų analizė. Taip pat čia apdorojamos ir interpretuojamos robots.txt faile pateiktos paieškos robotų taisyklės.



2.5 pav. Internetinės paieškos roboto architektūra

2.1.9. Problemos ir jų sprendimo būdai

Internetinio paieškos proceso projektavimo bei tolimesnio sistemos naudojimo metu dažnai iškyla įvairių problemų. Kad sukurtas paieškos robotas veiktų efektyviai, reikia šiuos keblumus išspręsti arba sumažinti juos sukeliančių sąlygų riziką. Žemiau pateikiamos dažniausiai pasitaikančios internetinio paieškos proceso problemos bei veikimo aktualijos, aprašytos literatūros šaltiniuose [20] [21] [22].

- **Serverių pasiekiamumo problemos**

  Serveriai nėra amžini, jie gali būti išjungiami keletui dienų, savaičių arba visam laikui. Internetinis paieškos procesas kartą aplankęs serveryje saugomus puslapius ir juos suindeksavęs, po tam tikro laiko vėl sugrįžta atnaujinti informacijos. Jeigu serveris neveikia, šie duomenys tampa nepasiekiami. Tokiu atveju internetinis voras turi nustatyti naują pakartotinio duomenų atnaujinimo laiko limitą. Jeigu vėl nepavykstą keletą kartų rasti pageidaujamų puslapių, internetinis voras visam laikui nutraukia apsilankymus šiame serveryje.

- **Serverio resursų išnaudojimo problema**

  Internetinis paieškos voras per daug dažnai lankydamas tam tikro serverio svetainės puslapius gali apkrauti šio serverio darbą. Taip pat nepateikdamas savo identifikacinės informacijos, voras gali sukelti įtarimą serverio administratoriams, kurie gali imtis atitinkamų veiksmų siekdami užblokuoti paieškos voro prieigą prie serverio resursų. Šią problemą galima išspręsti padidinant laiko tarpą, po kurio voras pakartotinai apsilankys svetainėje. Siekiant išspręsti voro identifikavimo problemą, voras serveriui siunčiamose užklausose turi pateikti identifikavimo duomenis: tai galėtų būti: kūrėjų elektroninis paštas, telefono numeris.

- **Ryšio tarp serverio ir voro nutrūkimas**

  Kartais gali pasitaikyti atvejų kai internetinis paieškos voras pradeda siųstis duomenis iš serveryje patalpintos svetainės, tačiau dėl ryšio problemų siuntimas nutrūksta. Kad internetinis voras nelauktų amžinai, kol siuntimas bus baigtas, kiekviena voro operacija privalo turėti tam tikrą operacijos vykdymo laiko limitą.

- **Puslapio antraštinės informacijos nagrinėjimas**

  Prieš siųsdamas į saugyklą duomenis, internetinis voras privalo patikrinti jų antraštinės informaciją. Ši informacija nurodo duomenų tipą ir leidžia vorui atmesti tuos duomenų rinkinius, kurių apdorojimui jis nėra pritaikytas.

- **Atsisiunčiamų duomenų kiekio ir dydžio apribojimas**

  Svarbu apriboti siunčiamų duomenų kiekį. Tai galima padaryti nustatant tam tikrą puslapių skaičių, kuriuos leidžiama atsisiųsti internetiniam vorui iš tos pačios svetainės. Taip pat galima riboti vieno atsisiunčiamo failo dydį.

- **Tų pačių puslapių atsisiuntimo problema**

  Kartais pasitaiko, kad internetinis paieškos robotas gali keletą kartų parsisiųsti tą patį puslapį. Išvengti šios problemos padeda parsisiųsto puslapio unikalus kodas. Šis kodas unikaliai identifikuoja puslapį ir leidžia kitąkart atmesti to pačio puslapio atsisiuntimą. Vis dėlto šis metodas turi ir trūkumų. Jis negali atskirti tokio atvejo, kai tam tikra puslapio duomenų dalis kartojasi.

- **Dinamiškų URL analizė**

  Svarbu atsižvelgti ir į puslapių dinaminių URL naudojimą. Jeigu puslapyje yra naudojamas dinamiškas parametras, pavyzdžiui sesijos ID, tas pats puslapis internetinio voro gali būti interpretuojamas kaip skirtingi puslapiai. Siekiant to išvengti internetinis paieškas procesas turi atpažinti, kad šis parametras neįtakoja turinio.

- **Spąstai internetiniams paieškos robotams**

  Kai kurie puslapiai gali būti pavojingi internetiniams paieškos robotams. Jeigu svetainė turi begalinio gylio vidinius puslapius arba ciklišKai sujungtus puslapius, internetinis paieškos procesas gali amžinai klaidžioti ir niekada nepasiekti paieškos pabaigos. Tokios problemos nėra pilnai išsprendžiamos, tačiau kai kurie spąstai paieškos vorams gali būti aptinkami ir išvengiami.

- **Tinkami klaidų pranešimai**

  Ši problema yra labiau susijusi su serverių nustatymais nei su internetinių paieškos vorų funkcionavimu. Jeigu serverio administratorius naudoja savo sukurtą „puslapis nerastas“ metodiką vartotojo informavimui apie neegzistuojančius puslapius, tai ši metodika gali būti netinkama internetiniams paieškos robotams. Siekiant išvengti šių nesusipratimų administratorius privalo tinkamai nustatyti serverį, kad jis gebėtų pateikti tinkamą informaciją apie nerastą puslapį ir paieškos robotams.

- **Puslapio turinio naujumo užtikrinimas**

  Naujienų portalai ir kiti interaktyvūs puslapiai dažnai atnaujina pateikiamą informaciją. Internetinis paieškos procesas tokias svetaines turi lankyti dažniau, tam, kad užtikrintų

duomenų „šviežumą“. Vis dėlto dėl per didelio atnaujinimų kiekio gali kilti keblumų, todėl internetinis paieškos procesas naudoja tam tikrą duomenų atnaujinimo laiko limitą bei puslapių svarbos nustatymą: pirmiausiai atnaujinami puslapiai, kurie yra svarbiausi. Puslapio prioritetas nustatomas pagal specialius kriterijus gaunamus taikant konkrečią paieškos strategiją.

2.2. Literatūros analizės išvados

Atliktos literatūros analizės metu gauti tokie rezultatai:

- Internetinę paiešką atlieka internetinis voras arba paieškos robotas.
- Pagrindinės internetinės paieškos proceso panaudojimo galimybės yra tokios: specializuotų katalogų sudarymas, autorinės apsaugos palaikymas, svetainių naujienų sekimas, svetainės kodo validavimas, svetainių turinio kopijavimas, specifinės informacijos paieška.
- Pagrindinės efektyvaus paieškos roboto savybės yra tokios: lankstumas, taupus serverio resursų naudojimas, našumas, patikimumas.
- Pagrindiniai paieškos procesų tipai yra šie: įprastiniai paieškos procesai, naujienų agentai, specializuotos paieškos robotai, svetainių kopijavimo sistemos, archyvų agentai, paslėpto interneto (Hidden web) paieškos sistemos.
- Nustatyta, kad didžiausios paieškos proceso problemos kyla dėl milžiniško interneto dydžio, nuolatinės informacijos kaitos ir svetainių dinamiškumo.

3. Projektinė dalis

Magistrinio baigiamojo darbo projektinėje dalyje pateikiama informacija apie sukurtą internetinės paieškos proceso modelį. Čia aprašoma realizuotos sistemos naudojimo paskirtis, diegimo aplinka, priimti esminiai sprendimai, apibrėžiamas sistemos veikimo kontekstas.

Šiame skyriuje taip pat suformuluojami sistemos funkciniai bei nefunkciniai reikalavimai, aprašoma sistemos vartotojų hierarchijos modelis, pateikiamas detalus sistemos teikiamų funkcijų (panaudojimo atvejų) sąrašas.

Šioje dalyje taip pat pateikiami esminiai sistemos architektūros sprendimai, aprašomas apibendrintas internetinio paieškos proceso architektūros modelis, detalizuojami pagrindiniai sistemos komponentai. Čia taip pat apibrėžiama sistemos naudojama duomenų bazės struktūra – duomenų modelis.

3.1. Sistemos paskirtis

Pagrindinis magistrinio darbo tikslas realizuoti internetinės paieškos procesą, skirtą informacijos internete paieškai bei rinkimui. Sukurtas paieškos robotas suteiks galimybę vartotojui vaizdžiai stebėti internetinės paieškos procesą, leis lanksčiai valdyti įvairius parametrus bei stebėti jų įtaką paieškos proceso efektyvumui.

Sistema kuriama ne komerciniais tikslais ir nebus platinama visuomenei. Internetinės paieškos robotas bus naudojamas užsakovo organizacijoje.

Ateityje, papildžius internetinio paieškos proceso funkcionalumą, jis galėtų būti panaudotas kaip sudėtinė dalis kuriant pilnavertę internetinės paieškos sistemą.

Magistrinio darbo metu sukurtos sistemos funkcionalumui buvo iškelti tokie esminiai tikslai:

- Internetinis paieškos procesas privalo gebėti efektyviai nagrinėti svetaines bei jų turinį. Sistemos funkcionalumas turi tenkinti šiuos reikalavimus: galimybė nustatyti ryšius tarp nagrinėtų svetainių, galimybė rasti svetainių vidines nuorodas bei įvertinti vidinių nuorodų svarbos koeficientus;
- Sistema vartotojui turi pateikti detalius rezultatus apie jau lankytas svetaines. Pateikiami tokie rezultatai: svetainės nuorodų skaičius, parsiųstų puslapių skaičius, vidutinis užkrovimo greitis,

svetainės meta duomenys ir t.t. Paieškos proceso darbo rezultatai turi būti pateikiami lentelėmis bei grafikais;

- Sukurta programinė sistema turi leisti vartotojui lanksčiai valdyti įvairius paieškos proceso parametrus;

Plačiau apie sistemos funkcionalumą galite skaityti 3.6 skyriuje „Panaudos atvejai“.

3.2. Esminiai projektavimo sprendimai

Projektavimo metu priimti šie esminiai sistemos realizacijos sprendimai:

- Internetinis paieškos procesas (robotas) realizuojamas PHP programavimo kalba, informacijos saugojimui naudojama MySQL duomenų bazių valdymo sistema;
- Sukurta sistema remiasi kliento – serverio architektūros pagrindu;
- Programinės sistemos komponentai realizuoti objektiškai, naudojami MVC (Model – View – Controller) architektūros principai;

3.3. Diegimo aplinka

Magistrinio darbo metu sukurta sistema – internetinis paieškos procesas – paprastai yra diegiamas internetiniame serveryje. Vis dėlto suteikiama galimybė šią sistemą naudoti ir asmeniniame kompiuteryje: papildomai reikia įdiegti virtualaus serverio programinę įrangą (pavyzdžiui: Xampp, Wamp). Žemiau pateikiami aplinkos reikalavimai, kai sistema diegiama internetiniame serveryje.

Serverio techninės bei programinės įrangos reikalavimai:

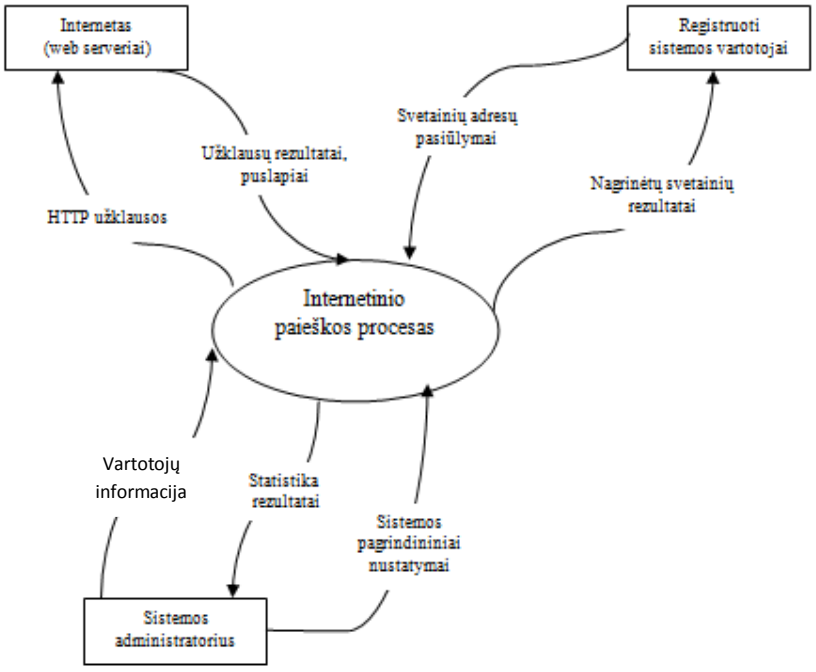
- Bent 64 megabaitai (MB) darbinės atminties (RAM);
- Mažiausiai 22 megabaitai serverio kietajame diske (HDD). Jeigu paieškos procese įjungta svetainių turinio parsisiuntimo galimybė, sistemai rekomenduojama paskirti dar keletą šimtų megabaitų papildomos vietos;
- Sistemos darbui užtikrinti reikalinga 5.2 arba naujesnė PHP versija. Rekomenduojama 5 arba naujesnė MySQL duomenų bazės versija;
- Serveryje, kuriame naudojamas paieškos robotas, turi būti įdiegta Linux pagrindu paremta operacinė sistema;
- Serveris turi palaikyti CronJob periodinių komandų vykdymą;

3.4. Veiklos kontekstas

Internetinio paieškos proceso veiklos sritis nėra labai plati. Ji yra sudaryta iš šių pagrindinių keturių dalyvių:

- Internetinis paieškos procesas. Pagrindinis veiklos konteksto dalyvis, kuris sąveikauja su kitais veiklos nariais (registruotais vartotojais, administratoriumi, internetiniais serveriais);
- Registruoti sistemos vartotojai. Tai vartotojai, kuriems administratorius suteikė galimybę naudotis sistemos paslaugomis. Registruoti vartotojai gali naudotis tik ribotu sistemos funkcionalumo paketu;
- Internetas (serveriai). Tai vienas svarbiausių veiklos konteksto dalyvių. Informacijos srautai tarp interneto bei internetinio paieškos proceso yra didžiausi visoje veikloje;
- Sistemos administratorius. Pagrindinis sistemos vartotojas, kontroliuojantis paieškos proceso darbą;

Žemiau pateikiama internetinio paieškos proceso veiklos konteksto diagrama bei pagrindinių veiklos įvykių lentelė.



3.1 pav. Internetinio paieškos proceso veiklos konteksto diagrama

Veiklos įvykių lentelėje pateikiama tokia informacija: įvykio pavadinimas bei įeinantys (in) ir išeinantys (out) informacijos srautai. Srauto tipas nustatomas internetinio paieškos proceso (žiūrėti aukščiau esantį paveikslą) atžvilgiu.

Pastaba: veiklos įvykių sąraše pateikiama tik dalis visų galimų įvykių.

3.1 lentelė. Veiklos įvykių sąrašas

Eil. Nr.	Įvykio pavadinimas	Įeinantys / išeinantys informacijos srautai
1.	Vartotojas paieškos robotui pasiūlo naują svetainės adresą	Svetainės adresas (in)
2.	Vartotojas pageidauja išsisaugoti nagrinėtos svetainės analizės informaciją	Svetainės adresas (in) Nagrinėtos svetainės rezultatai (out)
3.	Paieškos procesas patikrina ar tam tikra nuoroda egzistuoja	HTTP užklausa (out)
4.	Paieškos robotas iš internetinio serverio gauna užklaustos nuorodos puslapio turinį	HTTP užklausa (out) Puslapio turinys (in)
5.	Pakeičiami sisteminiai parametrai	Sistemos parametrų reikšmės (in)
6.	Administratorius peržiūri pageidaujamos svetainės statistikos informaciją	Svetainės adresas (in) Statistikos informacija (out)
7.	Užregistruojamas naujas sistemos vartotojas	Vartotojo informacija (in)
8.	Peržiūrimas parsiųstų svetainės puslapių sąrašas	Svetainės adresas (in) Puslapių sąrašas (out)

3.5. Sistemos vartotojai

Internetinio paieškos proceso vartotojai skirstomi į du tipus. Kiekvienam vartotojo tipui prieinamas apibrėžtas sistemos funkcionalumas.

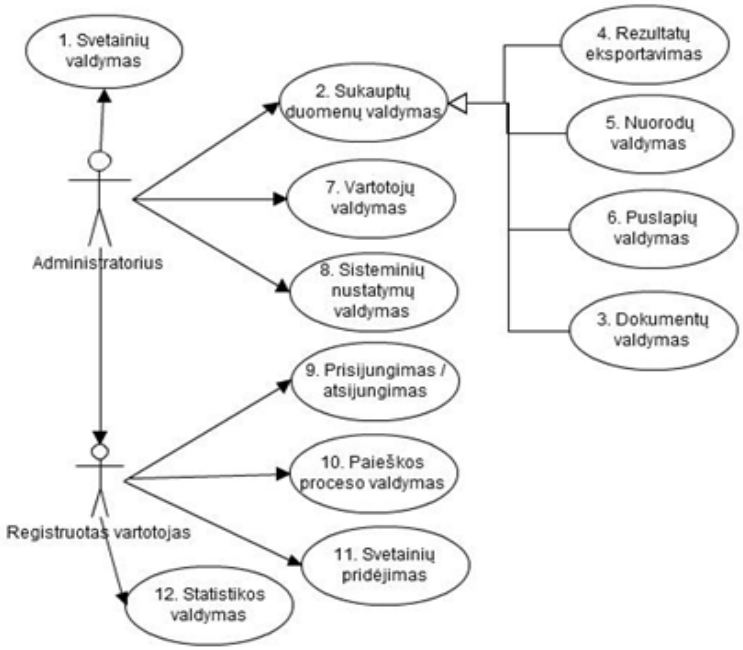
Sistemos vartotojų tipai:

- Sistemos administratorius. Aukščiausio lygio sistemos vartotojas, kuriam prieinamos visos sistemos teikiamos funkcijos;
- Registruotas vartotojas. Tai sistemos administratoriaus užregistruotas vartotojas, kuriam suteikiama galimybė pasiūlyti svetainių adresus, stebėti analizės rezultatus. Kiekvienas vartotojas gali naudotis ir kitomis sistemos paslaugomis, jeigu administratorius jam suteikė reikiamas teises.

Plačiau apie vartotojams prieinamas funkcijas galite skaityti kitame skyriuje 3.6 „Panaudojimo atvejai“.

3.6. Panaudojimo atvejai

Žemiau pateikiama sistemos panaudojimo atvejų (PA) diagrama bei detalizuotas visų PA sąrašas.



3.2 pav. Panaudos atvejų (PA) diagrama

Sistemos panaudojimo atvejų sąrašas:

- ✓ 1. Svetainių valdymas;
- ✓ 2. Sukauptų duomenų valdymas;
- ✓ 3. Dokumentų valdymas;
- ✓ 4. Rezultatų eksportavimas;
- ✓ 5. Nuorodų valdymas;
- ✓ 6. Puslapių valdymas;
- ✓ 7. Vartotojų valdymas;
- ✓ 8. Sisteminių nustatymų valdymas;
- ✓ 9. Prisijungimas / atsijungimas;
- ✓ 10. Paieškos proceso valdymas;
- ✓ 11. Svetainių pridėjimas;
- ✓ 12. Statistikos valdymas;

3.2 lentelė. Svetainių valdymo PA

1. PANAUDOJIMO ATVEJIS: Svetainių valdymas **Vartotojas/Aktorius:** Administratorius **Aprašas:** Užregistruotų svetainių valdymas **Panaudojimo atvejo scenarijus:** Paspaudžiamas pagrindinio meniu skyrius „Svetainės“; Pasirenkama svetainė; Pakoreguojama / peržiūrima svetainės informacija; **Prieš sąlyga:** Būtina prisijungti prie sistemos. Svetainių sąrašas negali būti tuščias. **Sužadinimo sąlyga:** iškilo poreikis peržiūrėti arba pakoreguoti svetainių informacija. **Po-sąlyga:** atliekami pasirinkti veiksmai, atnaujinama svetainės informacija.

3.3 lentelė. Sukauptų duomenų valdymo PA

2. PANAUDOJIMO ATVEJIS: Sukauptų duomenų valdymas **Vartotojas/Aktorius:** Administratorius **Aprašas:** Peržiūrimi arba koreguojami sukaupti duomenys. **Panaudojimo atvejo scenarijus:** Pasirenkamas pagrindinio meniu skyrius „Sukaupta informacija“; Vartotojas pasirenka duomenų tipą (svetainės, nuorodos, puslapiai); Pasirenkami elementai; Peržiūrima / pakoreguojama pasirinkto elemento informacija; **Prieš sąlyga:** Paieškos procesas turi būti įvykdytas bent kartą. Turi būti sukauptas tam tikras kiekis informacijos. **Sužadinimo sąlyga:** iškilo poreikis keisti / peržiūrėti paieškos proceso sukauptus duomenis. **Po-sąlyga:** peržiūrimi arba atnaujinami sukaupti duomenys.

3.4 lentelė. Dokumentų valdymo PA

3. PANAUDOJIMO ATVEJIS: Dokumentų valdymas
Vartotojas/Aktorius: Administratorius
Aprašas: Valdomos internetinio paieškos roboto sukauptos dokumentų nuorodos bei tipai.
Panaudojimo atvejo scenarijus: 1) Ieškomų dokumentų nustatymo atvejis Pasirenkamas meniu punktas „Sisteminiai nustatymas“; Pasirenkamas poskyris „Duomenų kaupimas“; Pakoreguojama / peržiūrima svetainės informacija; 2) Dokumentų nuorodų peržiūros atvejis Pasirenkamas pagrindinio meniu skyrius „Dokumentai“; Peržiūrimos arba pakoreguojamas dokumentų nuorodos;
Prieš sąlyga: Dokumentų peržiūros atveju, dokumentų sąrašas negali būti tuščias.
Sužadinimo sąlyga: iškilo poreikis ieškoti naujų dokumentų arba peržiūrėti sukauptų dokumentų informaciją.
Po-sąlyga: įvedamas naujas dokumento tipas, peržiūrima dokumentų informacija.

3.5 lentelė. Rezultatų eksportavimo PA

4. PANAUDOJIMO ATVEJIS: Rezultatų eksportavimas
Vartotojas/Aktorius: Administratorius
Aprašas: Eksportuojami pasirinkti paieškos proceso rezultatai.
Panaudojimo atvejo scenarijus: Pasirenkamas meniu skyrius „Statistika“; Pasirenkami duomenų filtravimo parametrai; Paspaudžiamas mygtukas „Eksportuoti duomenis“;
Prieš sąlyga: paieškos proceso metu turi būti išsaugotas bent minimalus kiekis informacijos.
Sužadinimo sąlyga: iškilo poreikis išsaugoti rezultatus išorinėje laikmenoje.
Po-sąlyga: paieškos rezultatai išsaugomi vartotojo laikmenoje.

3.6 lentelė. Nuorodų valdymo PA

5. PANAUDOJIMO ATVEJIS: Nuorodų valdymas **Vartotojas/Aktorius:** Administratorius **Aprašas:** Peržiūrima arba koreguojama sukaupta svetainių nuorodų informacija. **Panaudojimo atvejo scenarijus:** Pasirenkamas meniu skyrius „Nuorodos“; Peržiūrimos arba pakoreguojamos sukauptos svetainių nuorodos; **Prieš sąlyga:** paieškos procesas turi būti įvykdytas bent kartą; **Sužadinimo sąlyga:** iškilo poreikis peržiūrėti arba pakoreguoti sukauptą nuorodų informaciją. **Po-sąlyga:** peržiūrima (arba pakoreguojama) reikiama nuorodų informacija.

3.7 lentelė. Puslapių valdymo PA

6. PANAUDOJIMO ATVEJIS: Puslapių valdymas **Vartotojas/Aktorius:** Administratorius **Aprašas:** Peržiūrima arba koreguojama sukaupta puslapių turinio informacija. **Panaudojimo atvejo scenarijus:** Pasirenkamas meniu skyrius „Puslapiai“. Peržiūrimi arba pakoreguojami sukaupti svetainių puslapiai; **Prieš sąlyga:** puslapių sąrašas negali būti tuščias. Paieškos procesas turi būti paleistas bent kartą. **Sužadinimo sąlyga:** iškilo poreikis peržiūrėti sukauptą puslapių turinį. **Po-sąlyga:** peržiūrima arba pakoreguojama svetainių puslapio informacija.

3.8 lentelė. Vartotojų valdymo PA

7. PANAUDOJIMO ATVEJIS: Vartotojų valdymas **Vartotojas/Aktorius:** Administratorius **Aprašas:** Valdomi registruoti sistemos vartotojai. **Panaudojimo atvejo scenarijus:** Pasirenkamas meniu skyrius „Vartotojai“; Pasirenkamas konkretus vartotojas; Peržiūrima / pakoreguojama pasirinkto kliento informacija; **Prieš sąlyga:** turi būti įjungta registruotų vartotojų prieigos galimybė. **Sužadinimo sąlyga:** iškilo poreikis sukurti arba redaguoti vartotojų informaciją **Po-sąlyga:** atnaujinama vartotojų informacija

3.9 lentelė. Sisteminių nustatymų valdymo PA

8. PANAUDOJIMO ATVEJIS: Sisteminių nustatymų valdymas
Vartotojas/Aktorius: Administratorius
Aprašas: Sisteminių nustatymų ir paieškos proceso nustatymų valdymas.
Panaudojimo atvejo scenarijus: Pasirenkamas meniu skyrius „Sisteminiai nustatymai“; Pasirenkamas parametrų pogrupis; Atnaujinamos pasirinktų parametrų reikšmės;
Prieš sąlyga: būtina prisijungti prie administratoriaus valdymo pulto.
Sužadinimo sąlyga: iškilo poreikis pakoreguoti sisteminius parametrus.
Po-sąlyga: atnaujinami sisteminiai parametrai.

3.10 lentelė. Prisijungimo / atsijungimo PA

9. PANAUDOJIMO ATVEJIS: Prisijungimas / atsijungimas
Vartotojas/Aktorius: Administratorius, registruotas vartotojas
Aprašas: Prisijungiama arba atsijungiama nuo sistemos.
Panaudojimo atvejo scenarijus: **1)** Prisijungimo atveju: Iškviečiamas sistemos vartotojo prisijungimo puslapis; Užpildoma prisijungimo forma ir paspaudžiamas mygtukas „Prisijungti“; **2)** Atsijungimo atveju: Pasirenkamas meniu skyrius „Atsijungti“;
Prieš sąlyga: asmuo yra registruotas sistemoje bei turi prisijungimo duomenis.
Sužadinimo sąlyga: iškilo poreikis naudotis sistemos paslaugomis.
Po-sąlyga: asmuo prisijungia arba atsijungia nuo sistemos.

3.11 lentelė. Paieškos proceso valdymo PA

10. PANAUDOJIMO ATVEJIS: Paieškos proceso valdymas
Vartotojas/Aktorius: Administratorius.
Aprašas: Paleidžiamas arba sustabdomas paieškos procesas.
Panaudojimo atvejo scenarijus: Pasirenkamas meniu punktas „Paieškos procesas“; Paspaudžiamas reikiamas proceso valdymo mygtukas;
Prieš sąlyga: asmuo turi būti prisijungęs prie administratoriaus valdymo punkto.
Sužadinimo sąlyga: iškilo poreikis pakoreguoti paieškos proceso darbą.
Po-sąlyga: sustabdomas / įjungiamas paieškos procesas.

3.12 lentelė. Svetainių pridėjimo PA

11. PANAUDOJIMO ATVEJIS: Svetainių pridėjimas
Vartotojas/Aktorius: Administratorius, registruotas vartotojas
Aprašas: Pasiūlomas naujas svetainės adresas.
Panaudojimo atvejo scenarijus: Pasirenkamas meniu skyrius „Svetainės adreso pasiūlymas“; Įvedamas naujos svetainės adresas;
Prieš sąlyga: asmuo turi turėti teisę siūlyti svetainių adresus.
Sužadinimo sąlyga: iškilo poreikis į paieškos sistemą įtraukti naują svetainę.
Po-sąlyga: užregistruojama nauja svetainė.

3.13 lentelė. Statistikos valdymo PA

12. PANAUDOJIMO ATVEJIS: Statistikos valdymas
Vartotojas/Aktorius: Administratorius, registruotas vartotojas.
Aprašas: Peržiūrimi arba eksportuojami sukaupti paieškos proceso rezultatai.
Panaudojimo atvejo scenarijus: Pasirenkamas meniu skyrius „Statistika“. Užpildomi reikiami duomenų filtrai; Galima peržiūrėti / išeksportuoti paieškos kriterijus tenkinančius rezultatus.
Prieš sąlyga: asmuo turi turėti teisę peržiūrėti paieškos proceso statistikos rezultatus.
Sužadinimo sąlyga: iškilo poreikis peržiūrėti paieškos proceso darbo statistiką.
Po-sąlyga: peržiūrima dominanti paieškos statistikos informacija.

3.7. Funkciniai reikalavimai

Žemiau pateikiamas detalizuotas sistemai keliamų funkcinių reikalavimų sąrašas. Informacija pateikiama naudojantis Volere šablonais.

3.14 lentelė. Klaidų tikrinimo funkcinis reikalavimas

Reikalavimas #:	*1*	**Reikalavimo tipas:**	*1*	**Įvykis/panaudojimo atvejis #:**	*1, 2, 3, 5, 6 , 7, 8*
Aprašymas:	*Sistema turi informuoti vartotoją, kai jis suveda neteisingus duomenis.*				
Pagrindimas:	*Pateikti klaidingi duomenys gali stipriai sutrikdyti įprastinį sistemos darbą.*				
Šaltinis:	*Augustinas Gustas*				
Tikimo kriterijus:	*Sistema informuoja vartotoją, jei įvedami neteisingi duomenys.*				
Užsakovo tenkinimas:	*1*		**Užsakovo netenkinimas:**	*2*	
Priklausomybės:	Nėra		**Konfliktai:**	Nėra	
Papildoma medžiaga:	Nėra				
Istorija:	*Užregistruotas 2010 kovo 9 d.*				

3.15 lentelė. Klientų IP blokavimo funkcinis reikalavimas

Reikalavimas #:	*2*	**Reikalavimo tipas:**	*2*	**Įvykis/panaudojimo atvejis #:**	*7, 9*
Aprašymas:	*Sistema turi automatiškai blokuoti vartotojo IP adresą po tam tikro skaičiaus nesėkmingų prisijungimo bandymų. Sistemos administratoriui taip pat suteikiama galimybė rankiniu būdu blokuoti tam tikrus IP adresus.*				
Pagrindimas:	*IP blokavimas leidžia apriboti sistemos piktavališko naudojimo galimybes. Priėjimas prie sistemos suteikiamas tik siauram patikimų vartotojų ratui. IP blokavimas dažnai apsaugo nuo mėginimo įsilaužti į sistemą.*				
Šaltinis:	*Augustinas Gustas*				
Tikimo kriterijus:	*Užblokuojami piktavališki vartotojai.*				
Užsakovo tenkinimas:	*4*		**Užsakovo netenkinimas:**	*3*	
Priklausomybės:	Nėra		**Konfliktai:**	Nėra	

Papildoma medžiaga:	Nėra
Istorija:	***Užregistruotas 2010 kovo 9 d.***

3.16 lentelė. Rezultatų atvaizdavimo būdo pasirinkimo funkcinis reikalavimas

Reikalavimas #:	***3***	**Reikalavimo tipas:**	***3***	**Įvykis/panaudojimo atvejis #:**	***2, 12***
Aprašymas:	***Sistema turi leisti pasirinkti paieškos proceso darbo rezultatų atvaizdavimo būdą.***				
Pagrindimas:	***Sistemos darbo rezultatus reikia pateikti ne vien tekstine informacija, bet ir grafine. Šiuo atveju bus naudojami grafikai bei diagramos, kurie vaizdžiau pateiks rezultatus.***				
Šaltinis:	***Užsakovas***				
Tikimo kriterijus:	***Vartotojui suteikiama laisvė pasirinkti kaip peržiūrėti darbo rezultatus.***				
Užsakovo tenkinimas:	***5***		**Užsakovo netenkinimas:**	***4***	
Priklausomybės:	Nėra		**Konfliktai:**	Nėra	
Papildoma medžiaga:	Nėra				
Istorija:	***Užregistruotas 2010 kovo 10 d.***				

3.17 lentelė. Duomenų filtracijos funkcinis reikalavimas

Reikalavimas #:	***4***	**Reikalavimo tipas:**	***3***	**Įvykis/panaudojimo atvejis #:**	***1,2,3,4,5, 6, 12***
Aprašymas:	***Filtracijos parametrų pasirinkimas.***				
Pagrindimas:	***Vartotojui turi būti suteikiama galimybė filtruoti rezultatus pasirenkant tam tikrus kriterijus (pvz. datos apribojimus). Taip galima greičiau rasti arba išskirti tam tikrus darbo rezultatus.***				
Šaltinis:	***Užsakovas***				
Tikimo kriterijus:	***Išvedami rezultatai, tenkinantys pasirinktus kriterijus.***				
Užsakovo tenkinimas:	***4***		**Užsakovo netenkinimas:**		***3***
Priklausomybės:	Nėra		**Konfliktai:**	Nėra	
Papildoma medžiaga:	Nėra				
Istorija:	***Užregistruotas 2010 kovo 14 d.***				

3.18 lentelė. Darbo rezultatų spausdinimo funkcinis reikalavimas

Reikalavimas #:	*5*	**Reikalavimo tipas:**	*4*	**Įvykis/panaudojimo atvejis #:**	*2 , 12*
Aprašymas:	***Darbo rezultatų spausdinimas.***				
Pagrindimas:	***Vartotojui turi būti suteikiama galimybė atsispausdinti pasirinktus darbo rezultatus. Tai gali būti svarbu, kadangi tam tikrais atvejais reikia turėti spausdintinę rezultatų versiją.***				
Šaltinis:	***Augustinas Gustas***				
Tikimo kriterijus:	***Rezultatai pateikiami spausdintine forma.***				
Užsakovo tenkinimas:	*3*		**Užsakovo netenkinimas:**	*2*	
Priklausomybės:	Nėra		**Konfliktai:**	Nėra	
Papildoma medžiaga:	Nėra				
Istorija:	***Užregistruotas 2010 kovo 14 d.***				

3.19 lentelė. Dinamiško paieškos proceso valdymo funkcinis reikalavimas

Reikalavimas #:	*6*	**Reikalavimo tipas:**	*5*	**Įvykis/panaudojimo atvejis #:**	*10*
Aprašymas:	***Dinamiškas paieškos proceso reguliavimas.***				
Pagrindimas:	***Vartotojas turi gebėti dinamiškai valdyti paieškos proceso darbą. Esant poreikiui suteikiama galimybė įjungti, pristabdyti arba visai išjungti vykdomą darbą.***				
Šaltinis:	***Užsakovas***				
Tikimo kriterijus:	***Sistemos darbas gali būti valdomas bet metu.***				
Užsakovo tenkinimas:	*4*		**Užsakovo netenkinimas:**	*5*	
Priklausomybės:	Nėra		**Konfliktai:**	Nėra	
Papildoma medžiaga:	Nėra				
Istorija:	***Užregistruotas 2010 kovo 14 d.***				

3.20 lentelė. Paieškos nustatymų valdymo funkcinis reikalavimas

Reikalavimas #:	*7*	**Reikalavimo tipas:**	*6*	**Įvykis/panaudojimo atvejis #:**	*1, 5*
Aprašymas:	***Svetainės paieškos nustatymų keitimas.***				
Pagrindimas:	***Vartotojui suteikiama galimybė rankiniu būdu nustatyti tam tikrus paieškos parametrus konkrečiai svetainei. Tai pravartu, kai norima apriboti konkrečios svetainės puslapių adresų kiekį arba pakeisti informacijos atnaujinimo dažnumo intervalą.***				
Šaltinis:	**Užsakovas**				
Tikimo kriterijus:	***Kiekvienai svetainei nustatomi paieškos proceso parametrai.***				
Užsakovo tenkinimas:	*4*		**Užsakovo netenkinimas:**	*4*	
Priklausomybės:	Nėra		**Konfliktai:**	Nėra	
Papildoma medžiaga:	Nėra				
Istorija:	***Užregistruotas 2010 kovo 15 d.***				

3.21 lentelė. Svetainių blokavimo funkcinis reikalavimas

Reikalavimas #:	*8*	**Reikalavimo tipas:**	*7*	**Įvykis/panaudojimo atvejis #:**	*1, 5*
Aprašymas:	***Pasirinktų svetainių blokavimas.***				
Pagrindimas:	***Sistema turi leisti vartotojui užblokuoti paieškos proceso lankymąsi tam tikrose svetainėse. Užblokuojant kenksmingas svetaines galima išvengti neigiamos įtakos sistemos darbui (pvz. išvengiama paieškos robotų spąstų).***				
Šaltinis:	**Užsakovas**				
Tikimo kriterijus:	***Užblokuojamos pasirinktos svetainės / nuorodos.***				
Užsakovo tenkinimas:	*5*		**Užsakovo netenkinimas:**	*4*	
Priklausomybės:	Nėra		**Konfliktai:**	Nėra	
Papildoma medžiaga:	Nėra				
Istorija:	***Užregistruotas 2010 kovo 15 d.***				

3.22 lentelė. Sisteminių parametrų keitimo funkcinis reikalavimas

Reikalavimas #:	*9*	**Reikalavimo tipas:**	*6*	**Įvykis/panaudojimo atvejis #:**	*8*
Aprašymas:	***Sisteminių parametrų keitimas.***				
Pagrindimas:	***Vartotojas gali nustatyti bendrus sisteminius parametrus. Parametrai apriboja paieškos proceso darbą, sukauptų puslapių duomenų saugojimą.***				
Šaltinis:	**Užsakovas**				
Tikimo kriterijus:	***Nustatomi parametrai, įtakojantys sistemos darbą.***				
Užsakovo tenkinimas:	*5*	**Užsakovo netenkinimas:**	*5*		
Priklausomybės:	Nėra	**Konfliktai:**	Nėra		
Papildoma medžiaga:	Nėra				
Istorija:	***Užregistruotas 2010 kovo 15 d.***				

3.8. Nefunkciniai reikalavimai

Žemiau pateikiamas sąrašas nefunkcinių reikalavimų, apibūdinančių programinės sistemos išvaizdos, techninės dalies, duomenų saugojimo bei naudojimo charakteristikas. Nefunkciniai reikalavimai sudaryti bendraujant su sistemos užsakovu. Siekiant užtikrinti informacijos nuoseklumą reikalavimai pateikiami sugrupuotai pagal savo pobūdį.

Reikalavimai sistemos išvaizdai

Programinės įrangos vartotojo sąsajai keliami tokie reikalavimai:

- informatyvi, lengvai skaitoma bei suprantama sąsaja;
- intuityvus, nesudėtingas programos funkcijų valdymas;
- neįkyri sąsaja (nereikalaujama daug kartų patvirtinti kokį nors veiksmą);
- interaktyvi sąsaja. Programinė sistema interaktyviai sąveikauja su sistemos vartotoju;
- programinės sistemos vartotojo sąsaja tinkamai atvaizduojama esant įvairiai vartotojo monitoriaus rezoliucijai bei kitiems nustatymams.

Reikalavimai panaudojamumui

Programiniai sistemai keliami tokie panaudojamumo kriterijai:

- galimybė dirbti viena ranka. Tai leidžia sistema naudotis vartotojams su negalia;
- paprastai naudojamas asmenų, turinčių pagrindines kompiuterinio raštingumo žinias;
- sistemos pranešimai pateikiami nacionaline (lietuvių) kalba;
- paprasta naudotis IT srityje dirbantiems asmenims, kadangi pateikiami standartiniai srities trumpinimai bei žymėjimai;

Reikalavimai vykdymo charakteristikoms

Internetinio paieškos proceso veikimo charakteristikoms keliami tokie reikalavimai:

- greita ir talpi duomenų bazė (DB), gebėjimas efektyviai valdyti bei nagrinėti šimtus tūkstančių įrašų;
- taupus serverio resursų naudojimas. Sistemos darbas vykdomas optimaliai, netrikdant kitų tame pačiame serveryje arba kitur patalpintų internetinių sistemų darbo. Naudojamos priemonės efektyviai valdyti serverio procesoriaus apkrovimą, darbinės atminties užimtumą, duomenų srauto sunaudojimą;
- patikimas sistemos veikimas esant įvairiems serverių atsakymams;
- esama sistemos greitaveika užtikrina internetinės paieškos proceso efektyvų darbą. Paieškos procesas geba apdoroti priimtiną nuorodų kiekį per numatytą laiko intervalą;
- programinės sistemos gebėjimas greitai sureaguoti į vartotojo atliktus veiksmus;

Reikalavimai veikimo sąlygoms

Internetinio paieškos roboto veikimo sąlygoms keliami tokie reikalavimai:

- neaukšta oro temperatūra (būtina užtikrinti tinkamą serverio, kuriame įdiegta sistema, darbinės aplinkos vėdinimą);
- programinės sistemos naudojimui būtinas kompiuteris su įdiegta Javascript bei Ajax technologijas palaikančia internetine naršykle;

Reikalavimai sistemos priežiūrai

Reikalavimai programinės sistemos priežiūrai yra tokie:

- programinės įrangos produkto naudojimas bei palaikymas neturi reikalauti didelių išlaidų bei žmogiškųjų resursų poreikio;

- pasikeitus sistemą naudojančiai organizacijai ar organizacijos gyvavimo aplinkai neturi kilti didelių problemų atliekant reikiamus pakeitimus;
- ateityje sistema turi būti pritaikyta naudojimui serveriuose, su įdiegtomis kitomis operacinėmis sistemomis. Dabartinė palaikoma operacinė sistema – Unix;

Reikalavimai saugumui

Išskiriami tokie trys pagrindiniai programiniai sistemai keliami saugumo reikalavimai:

- konfidencialumas – sistemoje esantys duomenys apsaugoti nuo neteisėtos prieigos;
- vientisumas – sistemos duomenys vienareikšmiškai atitinka šaltinio perduotus (iš jo gautus) duomenis, kartu užtikrinant jų panaudojimo teisėtumą;
- pasiekiamumas – galimybė pasinaudoti sistemos duomenimis per fiksuotą laiko tarpą reikiamą prieigą turintiems vartotojams;

Kultūriniai-politiniai reikalavimai

Programiniai sistemai keliami tokie kultūriniai-politiniai reikalavimai:

- pagrindiniai sistemos komponentai kuriami Lietuvoje;
- esant poreikiui bus naudojamasi tik patikimų užsienio šaltinių komponentais;
- sistemoje nebus naudojama ką nors įžeidžiančių terminų ar iliustracijų;

Teisiniai reikalavimai

Internetiniam paieškos procesui keliami tokie teisiniai reikalavimai:

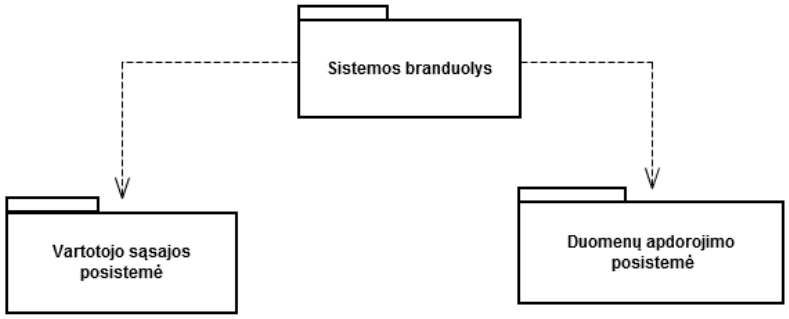
- Internetinis paieškos procesas negali neteisėtai kaupti konfidencialios kitų asmenų informacijos be jų sutikimo;
- Draudžiama naudoti surinktą informaciją rinkodaros arba kitais komercinę naudą suteikiančiais tikslais;

3.9. Architektūros specifikacija

3.9.1. Bendras sistemos architektūros modelis

Aukščiausio lygio paieškos proceso architektūros modelis susideda iš trijų posistemių. Žemiau pateikiami pagrindinių internetinio paieškos roboto posistemių aprašymai bei sistemos architektūros modelis aukščiausiame lygyje.

- ✓ Sistemos branduolys. Tai pagrindinis sistemos komponentas, apjungiantis informacijos apdorojimo bei vartotojo sąsajos generavimo posistemes;
- ✓ Vartotojo sąsajos posistemė. Ši posistemė atsakinga už vartotojo grafinės sąsajos generavimą, vartotojo atliktų veiksmų apdorojimą;
- ✓ Duomenų aprojimo posistemė. Ši posistemė apima komponentus, skirtus paieškos proceso metu gautos informacijos apdorojimui bei ilgalaikiam saugojimui duomenų bazėje;



3.3 pav. Aukščiausio lygio sistemos architektūros modelis

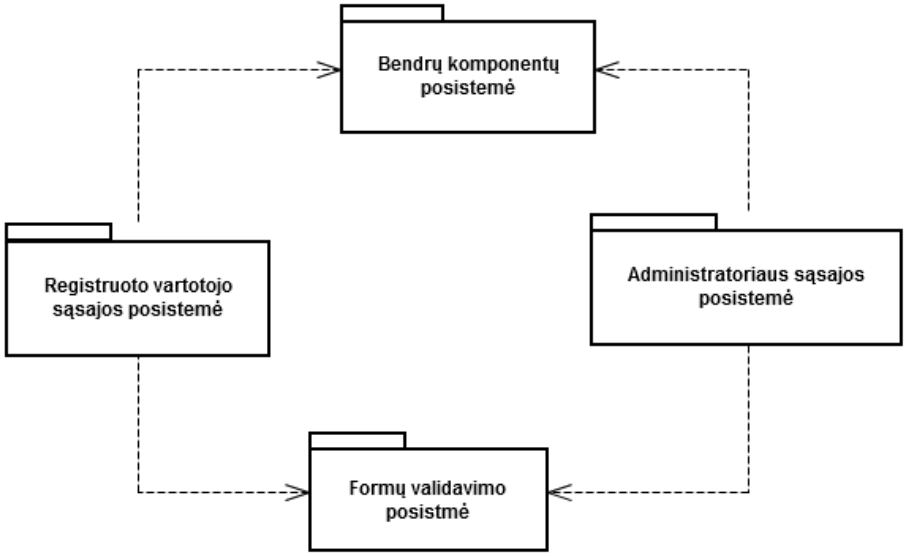
3.9.2. Detalesnis sistemos architektūros komponentų vaizdas

Žemiau pateikiamos detalizuotos aukščiausio lygio sistemos architektūros modelio komponentų diagramos.

Vartotojo sąsajos posistemės detalizuotas struktūros modelis

Vartojo sąsajos posistemės struktūrinis modelis susideda iš šių komponentų:

- ✓ Bendrų komponentų posistemė. Tai posistemė, kurioje apjungiamos klasės bendrai naudojamos registruotų bei administratoriaus sąsajų posistemių;
- ✓ Registruoto vartotojo sąsajos posistemė. Ši posistemė yra atsakinga už registruoto vartotojo sąsajos generavimą bei veiksmų apdorojimą;
- ✓ Administratoriaus sąsajos posistemė. Ši posistemė yra atsakinga už sistemos administratoriaus sąsajos generavimą bei veiksmų apdorojimą;
- ✓ Formų validavimo posistemė. Ši posistemė saugo klases, palengvinančias sąsajos formų validavimą;

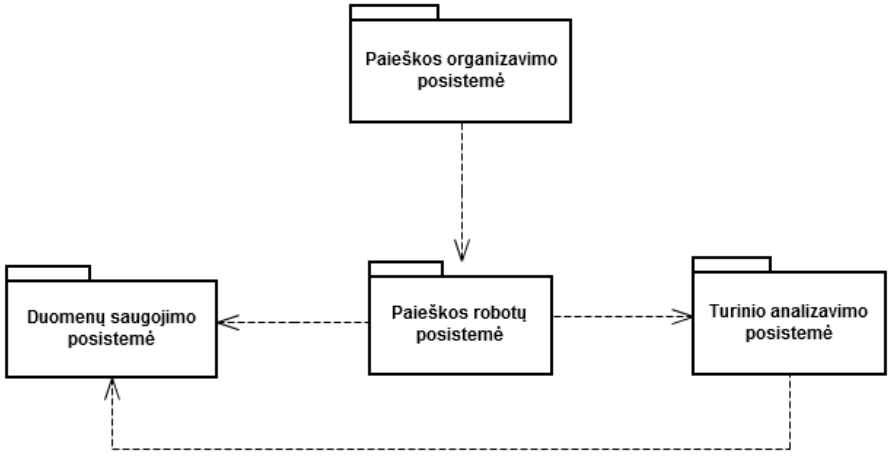


3.4 pav. Detalesnis vartotojo sąsajos posistemės architektūros modelis

Duomenų apdorojimo posistemės detalizuotas struktūros modelis

Duomenų apdorojimo posistemės struktūrinis modelis susideda iš šių komponentų:

- ✓ Paieškos organizavimo posistemė. Tai posistemė, apjungianti klases, skirtas paieškos proceso organizavimui. Čia nustatoma, kokios svetainių nuorodos bus analizuojamos pirmiau, kokios vėliau;
- ✓ Duomenų saugojimo posistemė. Ši posistemė atsakinga už informacijos duomenų bazėje išsaugojimą bei išrinkimą;
- ✓ Paieškos robotų posistemė. Robotų posistemė saugo pagrindines klases, reikalingas organizuoti paieškos robotų darbui;
- ✓ Turinio analizavimo posistemė. Šios posistemės klasės atlieka atsisiųsto puslapio turinio analizės funkciją. Čia išrenkamos svetainių, failų nuorodos;

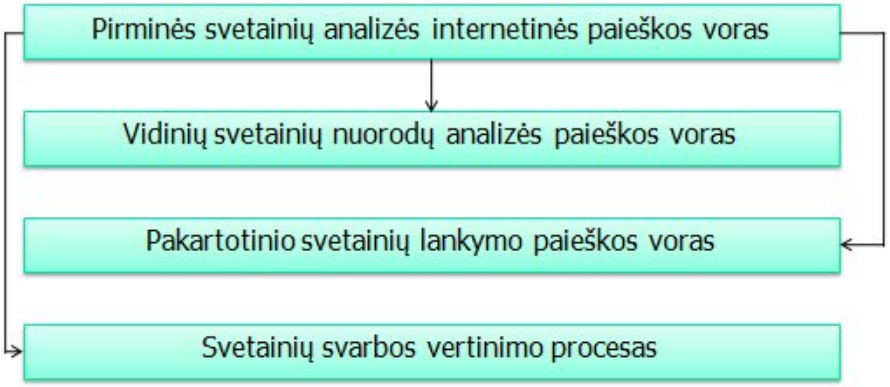


3.5 pav. Detalizuotas duomenų apdorojimo posistemės architektūros modelis

3.9.3. Paieškos proceso struktūra

Bendras internetinis paieškos procesas yra skaidomas į 4 žemesnio lygio sudedamąsias dalis (komponentus). Šie komponentai yra:

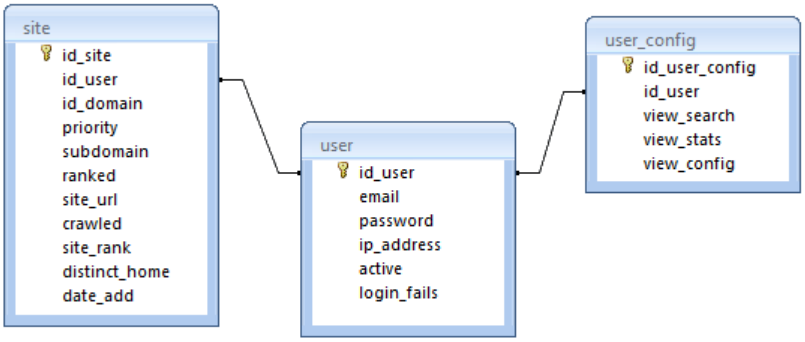
- ✓ Pirminės svetainių analizės internetinis paieškos voras. Kiekviena naujai surasta ar užregistruota svetainė pirmiausiai yra aplankoma pirminės svetainių analizės paieškos roboto. Pirmojo apsilankymo metu surandamos svetainės pagrindinės vidinės nuorodos, išrenkamos nuorodos, vedančios į kitas svetaines;
- ✓ Vidinių svetainių nuorodų analizės paieškos voras. Po pirminės analizės paieškos roboto apsilankymo svetainės analizei yra iškviečiamas vidinių svetainės nuorodų analizės paieškos voras. Šio paieškos roboto paskirtis – visų svetainės vidinių nuorodų suradimas;
- ✓ Pakartotinio svetainių lankymo paieškos voras. Tai sistema skirta pakartotiniam svetainių lankymui siekiant aptikti pasikeitusį svetainės turinį bei naujas nuorodas;
- ✓ Svetainių svarbos vertinimo procesas. Šis procesas periodiškai atnaujina užregistruotų svetainių bei nuorodų svarbos koeficientų reikšmes;



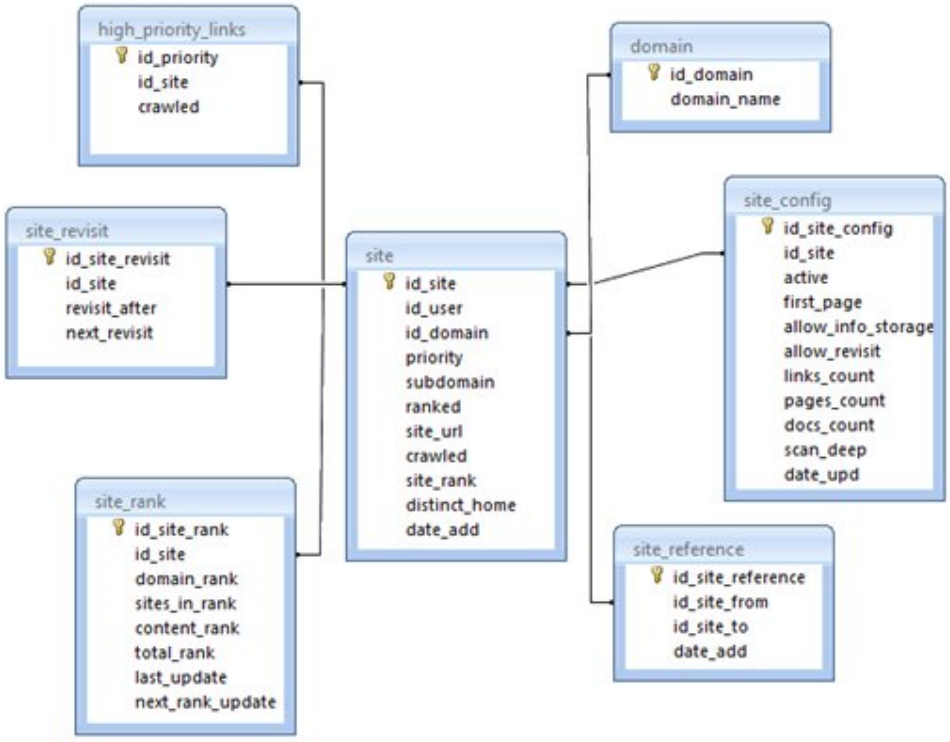
3.6 pav. Paieškos proceso skaidymas į žemesnio lygio komponentus

3.9.4. Duomenų bazės struktūros modelis

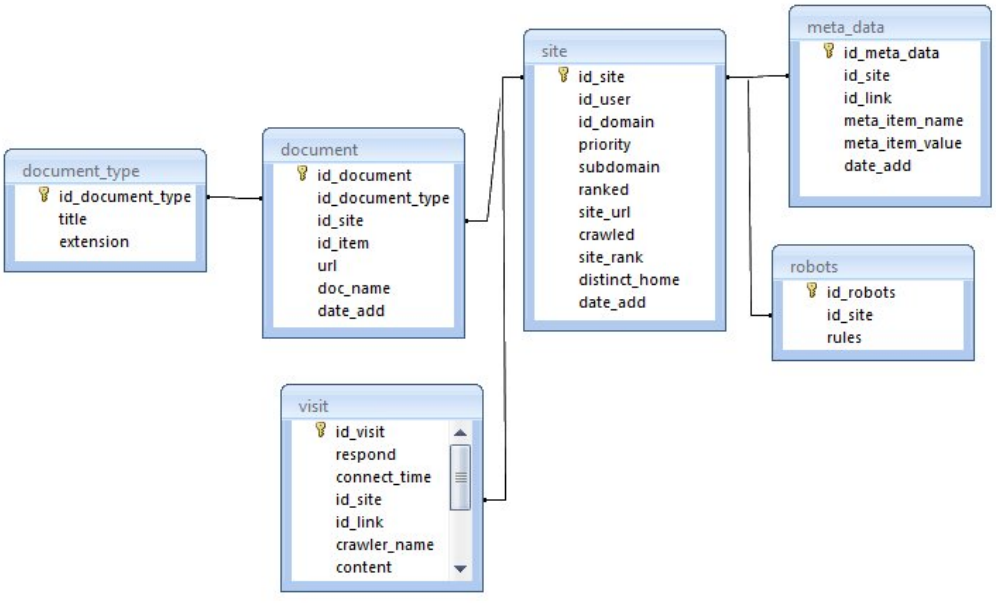
Žemiau pateikiamuose 3.7, 3.8 ir 3.9 paveiksluose atvaizduojamas magistrinio projekto sistemos duomenų bazės struktūros modelis. Siekdamas palengvinti diagramų skaitomumą, duomenų bazės struktūros bendrą modelį išskaidžiau į keletą mažesnių diagramų.



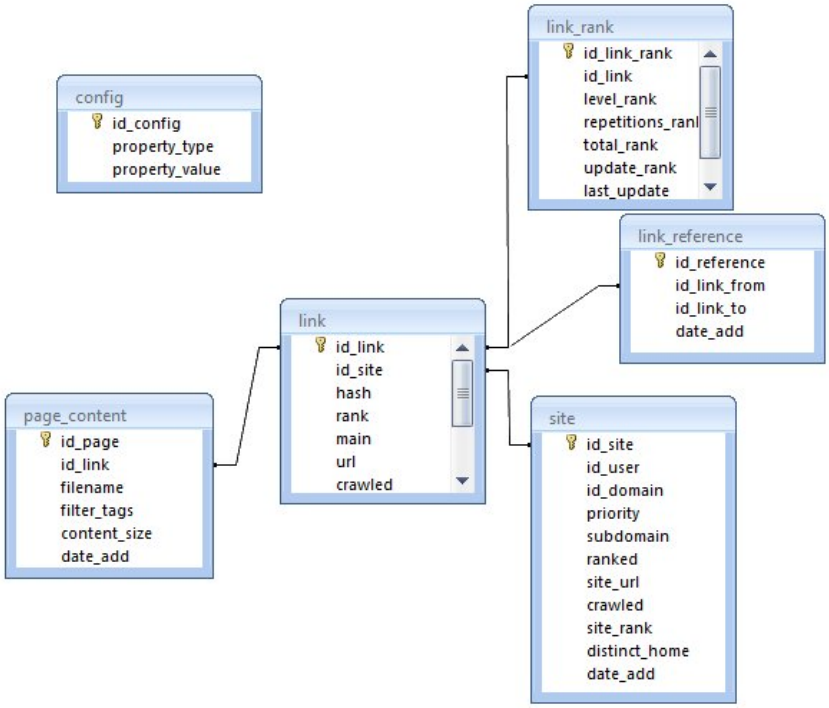
3.7 pav. Duomenų bazės lentelių struktūros modelis



3.8 pav. Duomenų bazės lentelių struktūros modelis



3.9 pav. Duomenų bazės lentelių struktūros modelis



3.10 pav. Duomenų bazės lentelių struktūros modelis

Žemiau esančiose lentelėse pateikiamos duomenų bazės lentelių specifikacijos.

3.23 lentelė. Duomenų bazės lentelės Site specifikacija

Pavadinimas	Site
Apibrėžimas	Svetainės informacija
Atsakomybės	Saugoma pagrindinės svetainių informacija
Sąveikavimas	Sąveikauja su lentele links,
Sąsaja/eksportas	Duomenų bazės lentelė neatlieka jokių veiksmų tik duomenų saugojimą

3.24 lentelė. Duomenų bazės lentelės User specifikacija

Pavadinimas	User
Apibrėžimas	Vartotojų informacija
Atsakomybės	Saugoma sistemos vartotojų informacija
Sąveikavimas	Sąveikauja su lentelėmis Site ir User_config
Sąsaja/eksportas	Duomenų bazės lentelė neatlieka jokių veiksmų tik duomenų saugojimą

3.25 lentelė. Duomenų bazės lentelės User_config specifikacija

Pavadinimas	User_config
Apibrėžimas	Vartotojų leidimų informacija
Atsakomybės	Saugoma sistemos vartotojų leidimų informacija
Sąveikavimas	Sąveikauja su lentele User
Sąsaja/eksportas	Duomenų bazės lentelė neatlieka jokių veiksmų tik duomenų saugojimą

3.26 lentelė. Duomenų bazės lentelės High_priority_links specifikacija

Pavadinimas	High_priority_links
Apibrėžimas	Aukštos svarbos svetainių sąrašas
Atsakomybės	Saugomas aukštos svarbos svetainių sąrašas. Į šį sąrašą patenka tos svetainės, kurios yra iškeliamos vartotojų.
Sąveikavimas	Sąveikauja su lentele Site
Sąsaja/eksportas	Duomenų bazės lentelė neatlieka jokių veiksmų tik duomenų saugojimą

3.27 lentelė. Duomenų bazės lentelės Site_revisit specifikacija

Pavadinimas	Site_revisit
Apibrėžimas	Svetainių pakartotinis lankymas
Atsakomybės	Saugoma informacija apie svetainių pakartotinį aplankymą.
Sąveikavimas	Sąveikauja su lentele Site
Sąsaja/eksportas	Duomenų bazės lentelė neatlieka jokių veiksmų tik duomenų saugojimą

3.28 lentelė. Duomenų bazės lentelės Site_rank specifikacija

Pavadinimas	Site_rank
Apibrėžimas	Svetainių svarbos koeficientai
Atsakomybės	Saugoma informacija apie nustatytus svetainių svarbos koeficientus
Sąveikavimas	Sąveikauja su lentele Site
Sąsaja/eksportas	Duomenų bazės lentelė neatlieka jokių veiksmų tik duomenų saugojimą

3.29 lentelė. Duomenų bazės lentelės Domain specifikacija

Pavadinimas	Domain
Apibrėžimas	Domenų informacija
Atsakomybės	Saugoma informacija apie svetainių domenų tipus
Sąveikavimas	Sąveikauja su lentele Site
Sąsaja/eksportas	Duomenų bazės lentelė neatlieka jokių veiksmų tik duomenų saugojimą

3.30 lentelė. Duomenų bazės lentelės Site_config specifikacija

Pavadinimas	Site_config
Apibrėžimas	Svetainių paieškos parametrai
Atsakomybės	Saugoma informacija apie nustatytus svetainių paieškos parametrus
Sąveikavimas	Sąveikauja su lentele Site
Sąsaja/eksportas	Duomenų bazės lentelė neatlieka jokių veiksmų tik duomenų saugojimą

3.31 lentelė. Duomenų bazės lentelės Site_reference specifikacija

Pavadinimas	Site_reference1
Apibrėžimas	Ryšiai tarp atskirų svetainių
Atsakomybės	Saugoma informacija apie ryšius tarp skirtingų svetainių.
Sąveikavimas	Sąveikauja su lentele Site
Sąsaja/eksportas	Duomenų bazės lentelė neatlieka jokių veiksmų tik duomenų saugojimą

3.32 lentelė. Duomenų bazės lentelės Document_type specifikacija

Pavadinimas	Document_type
Apibrėžimas	Ieškomų dokumentų (failų) tipai
Atsakomybės	Saugoma informacija apie vartotojo ieškomų failų tipus
Sąveikavimas	Sąveikauja su lentele Document
Sąsaja/eksportas	Duomenų bazės lentelė neatlieka jokių veiksmų tik duomenų saugojimą

3.32 lentelė. Duomenų bazės lentelės Document specifikacija

Pavadinimas	Document
Apibrėžimas	Rastų dokumentų sąrašas
Atsakomybės	Saugomas rastų dokumentų nuorodų sąrašas
Sąveikavimas	Sąveikauja su lentelėmis Site, Document_type
Sąsaja/eksportas	Duomenų bazės lentelė neatlieka jokių veiksmų tik duomenų saugojimą

3.33 lentelė. Duomenų bazės lentelės Visit specifikacija

Pavadinimas	Visit
Apibrėžimas	Apsilankymai
Atsakomybės	Saugoma informacija internetinių paieškos procesų apsilankymus svetainėse.
Sąveikavimas	Sąveikauja su lentele Site
Sąsaja/eksportas	Duomenų bazės lentelė neatlieka jokių veiksmų tik duomenų saugojimą

3.34 lentelė. Duomenų bazės lentelės Meta_data specifikacija

Pavadinimas	Meta_data
Apibrėžimas	Svetainių meta duomenys
Atsakomybės	Saugomi duomenys apie svetainių / nuorodų meta informaciją.
Sąveikavimas	Sąveikauja su lentelėmis Site, Link
Sąsaja/eksportas	Duomenų bazės lentelė neatlieka jokių veiksmų tik duomenų saugojimą

3.35 lentelė. Duomenų bazės lentelės Robots specifikacija

Pavadinimas	Robots
Apibrėžimas	Paieškos robotų taisyklės
Atsakomybės	Saugoma svetainės robots.txt faile esanti informacija
Sąveikavimas	Sąveikauja su lentele Site
Sąsaja/eksportas	Duomenų bazės lentelė neatlieka jokių veiksmų tik duomenų saugojimą

3.36 lentelė. Duomenų bazės lentelės Config specifikacija

Pavadinimas	Config
Apibrėžimas	Sistemos nustatymai
Atsakomybės	Saugomos sisteminių parametrų reikšmės
Sąveikavimas	Nėra
Sąsaja/eksportas	Duomenų bazės lentelė neatlieka jokių veiksmų tik duomenų saugojimą

3.37 lentelė. Duomenų bazės lentelės Page_content specifikacija

Pavadinimas	Page_content
Apibrėžimas	Puslapių turinys
Atsakomybės	Saugoma svetainės puslapių turinio informacija
Sąveikavimas	Sąveikauja su lentele Link
Sąsaja/eksportas	Duomenų bazės lentelė neatlieka jokių veiksmų tik duomenų saugojimą

3.38 lentelė. Duomenų bazės lentelės Link specifikacija

Pavadinimas	Link
Apibrėžimas	Svetainių nuorodos
Atsakomybės	Saugomas svetainių vidinių nuorodų sąrašas
Sąveikavimas	Sąveikauja su lentelėmis Site, Page_content, Link_rank, Link_reference
Sąsaja/eksportas	Duomenų bazės lentelė neatlieka jokių veiksmų tik duomenų saugojimą

3.39 lentelė. Duomenų bazės lentelės Link_rank specifikacija

Pavadinimas	Link_rank
Apibrėžimas	Nuorodų svarbos parametrai
Atsakomybės	Lentelėje saugomos svetainių vidinių nuorodų svarbos koeficientų reikšmės
Sąveikavimas	Sąveikauja su lentele Link
Sąsaja/eksportas	Duomenų bazės lentelė neatlieka jokių veiksmų tik duomenų saugojimą

3.40 lentelė. Duomenų bazės lentelės Link_reference specifikacija

Pavadinimas	Link_reference
Apibrėžimas	Nuorodų ryšiai
Atsakomybės	Saugoma informacija apie skirtingų svetainių nuorodų tarpusavio ryšius
Sąveikavimas	Sąveikauja su lentele Link
Sąsaja/eksportas	Duomenų bazės lentelė neatlieka jokių veiksmų tik duomenų saugojimą

3.10. Sistemos testavimas

Programinė įranga yra gera, jei ji atlieka numatytus veiksmus teisingai bei grąžina rezultatus per priimtiną laiko tarpą. Niekada negalima teigti, kad sukurta programinė įranga yra visu šimtu procentų ištestuota ir ,kad joje nėra klaidų. Testavimo tikslas – kiek galima daugiau sumažinti programinėje įrangoje esančių klaidų skaičių. Siekiant šio rezultato testavimo metu dirba aukštos kvalifikacijos testavimo komanda, naudojami patikimi testavimo įrankiai, parengiami unikalūs tik tai sistemai skirti testiniai atvejai, naudojama detali sistemos testavimo specifikacija.

Magistrinio darbo sistemos testavimui buvo naudojamas automatizuotas PHP testavimo įrankis „SimpleTest“. Testavimo metu ištestuoti komponentai, aprašyti aukščiau esančiame testavimo plane. Ištestuoti ne tik esminiai tų komponentų metodai (būtini sistemos funkcionavimui), bet ir pagalbiniai tų komponentų metodai.

Sistemos testavimo metu buvo atlikti šie testavimo etapai:

- ✓ Vienetų testavimas;
- ✓ Komponentų integracinis testavimas;
- ✓ Bendras sistemos funkcionavimo testavimas;

Testavimo rezultatai buvo kaupiami MS Word dokumentuose. Juose saugoma tokia informacija: testavimo atvejo aprašymas, testo pradiniai duomenys, laukiami ir gauti rezultatai.

3.11. Projektavimo dalies išvados

- Internetinis paieškos procesas realizuotas viena populiariausių internete naudojamų programavimo kalbų PHP.
- Projektuojant sistemą remtasi MVC architektūros principais, kadangi tai leido atskirti duomenų apdorojimo bei atvaizdavimo sluoksnius.

- Sistemos komponentai yra suskirstyti į šias pagrindines posistemes: sistemos branduolys, vartotojo sąsajos posistemė, duomenų apdorojimo posistemė. Vartotojo sąsajos posistemė yra sudaryta iš bendrų komponentų posistemės, formų validavimo posistemės, registruoto vartotojo sąsajos posistemės ir administratoriaus sąsajos posistemės. Duomenų apdorojimo posistemė apjungia paieškos organizavimo, duomenų saugojimo, paieškos robotų, turinio analizės posistemes. Toks sistemos komponentų jungimas į posistemes leido sugrupuoti klases pagal jų atliekamą funkcionalumą, kaip pvz.: administratoriaus sąsajos posistemė atsako už administratoriaus sąsajos generavimą bei valdymą.
- Sistemos paieškos procesas yra išskaidytas į šiuos žemesnio lygio komponentus: pirminės svetainės analizės paieškos voras, vidinių nuorodų analizės voras, pakartotinio svetainių lankymo voras, svetainių nuorodų svarbos vertinimo procesas. Toks skaidymas leido supaprastinti paieškos proceso struktūrą bei padalino paieškos procesą į atskirus komponentus, atliekančius konkrečius specifinius veiksmus.
- Sistemos testavimui buvo panaudotas automatizuoto testavimo įrankis „SimpleTest“. Šis įrankis leido automatizuoti testinių atvejų vykdymą bei rezultatų tikrinimą, kas leidžia užtikrinti didesnę testavimo kokybę bei patikimumą.

4. Tiriamoji dalis

4.1. Problemos aprašymas

Svarbi internetinio paieškos proceso organizavimo problema – svetainių svarbos koeficientų (prioritetų) nustatymas. Optimalus prioritetų nustatymas leidžia paieškos procesui pasirinkti kurioms svetainėms skirti didesnį dėmesį ir greičiau atlikti jų analizę. Svetainių svarbos nustatymas ypač aktualus organizuojant pakartotinį svetainių aplankymą bei svetainių vidinių nuorodų analizę.

4.2. Tyrimo aprašymas

Magistrinio darbo tyrimo metu nagrinėjamas paieškos proceso efektyvumo pasikeitimas naudojant svetainių svarbos metrikas. Taip pat analizuojama, kurioms iš pasirinktų metrikų priskirti didžiausią svorinį koeficientą skaičiuojant galutinę svetainių svarbos reikšmę.

Tyrimo metu naudojamas internetinis paieškos robotas, sukurtas baigiamojo magistrinio darbo metu. Tyrinėjamos dvi šios sistemos versijos. Vienoje sistemos versijoje svetainių paieška vykdoma iš eilės nagrinėjant visas svetaines, nėra vertinama svetainių svarba. Kitoje versijoje naudojamas svetainių prioritetų nustatymo algoritmas. Siekiama ištirti kaip patobulinimai įtakojo paieškos proceso efektyvumą.

Tyrimo metu naudojamas apibrėžtas internetinių svetainių rinkinys. Tyrimui pasirinkta 10 įvairių skirtingo pobūdžio svetainių. Svetainių sąrašą galite matyti žemiau esančioje 4.1 lentelėje. Dėl tolimesnio rezultatų palyginimo, lentelėje taip pat pateikiamos paieškos sistemos Google nustatytos svetainių svarbos reikšmės (angl. pagerank PR). Google PR reikšmės pateiktos remiantis svetainės http://www.prchecker.info/ 2011-05-19 dienos informacija.

4.1 lentelė. Tyrime naudojamų svetainių sąrašas

Nr.	Svetainės adresas	Svetainės aprašymas	Google PR
1.	http://www.delfi.lt	Vienas didžiausių Lietuvoje informacinių – naujienų portalų	7
2.	http://www.skelbiu.lt	Didžiausias Lietuvoje skelbimų portalas	5
3.	http://www.alfa.lt	Naujienų portalas	6
4.	http://www.pazintys.lt	Pažinčių portalas	5
5.	http://www.krepsinis.net/	Informacinis krepšinio naujienų portalas	6

Nr.	Svetainės adresas	Svetainės aprašymas	Google PR
6.	http://www.autogidas.lt	Automobilių pardavimo skelbimų portalas	5
7.	http://www.musulaikas.com	Regioninė informacinė svetainė	5
8.	http://www.microsoft.com	Microsoft korporacijos oficiali svetainė	9
9.	http://pigu.lt	Viena didžiausių Lietuvoje internetinė parduotuvė	6
10.	http://www.ktu.lt/	Kauno Technologijos universiteto oficiali svetainė	8

Žemiau pateikiamas pasirinktų svetainės svarbos skaičiavimo metrikų sąrašas. Dėl patogumo toliau naudojamos šių metrikų santrumpos, sudarytos iš metrikos pavadinime esančių žodžių pirmųjų raidžių.

Visas metrikas pagal pobūdį galima suskirstyti į šias grupes: svetainių tarpusavio ryšių metrikos (metrikos ĮSNS ir ISNS), vidinių nuorodų metrikos (VNSPS), domeno tipo metrikos (DT ir DG), informacijos atnaujinimo metrikos (NNS).

4.2 lentelė. Svetainių svarbos vertinimo metrikos

Nr.	Metrikos pavadinimas (santrumpa)	Metrikos aprašymas
1.	Įeinančių svetainių nuorodų skaičius (ĮSNS)	Tai skaičius svetainių, kurios turi nagrinėjamos svetainės nuorodą.
2.	Išeinančių svetainių nuorodų skaičius (ISNS)	Tai nagrinėjamoje svetainėje esančių kitų svetainių nuorodų skaičius.
3.	Vidinių nuorodų skaičius pradiniame puslapyje (VNSPS)	Tai pirmame nagrinėjamos svetainės puslapyje esančių vidinių nuorodų skaičius.
4.	Adreso galūnė (DG)	Nagrinėjamos svetainės URL adreso galūnė (domeno tipas).
5.	Naujų nuorodų skaičius (NNS)	Rastų naujų nuorodų skaičius pakartotinai aplankius svetainę po nustatyto apibrėžto laiko tarpo.

5. Eksperimentinė dalis

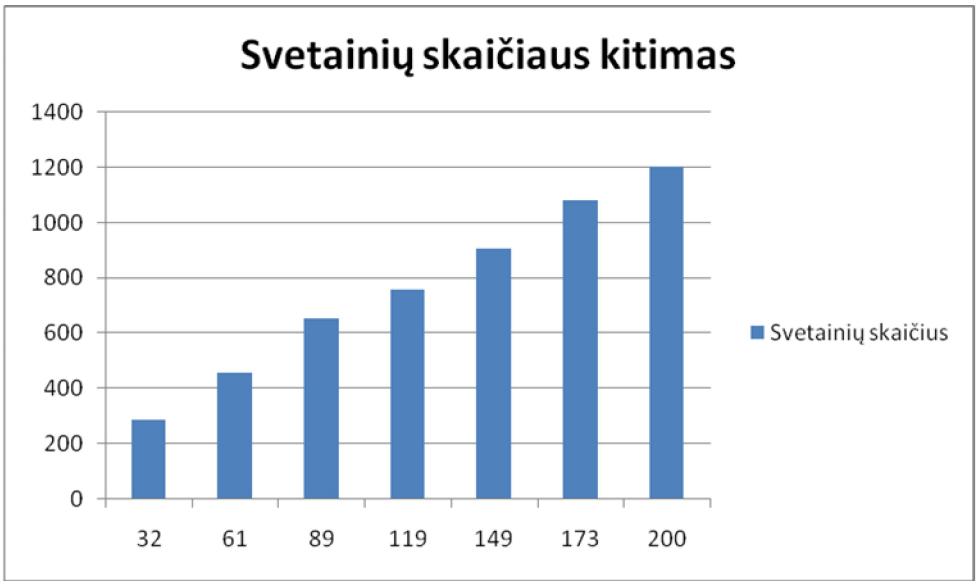
5.1. Pasiruošimas eksperimentui

Paieškos proceso svarbos parametrų tyrimo metu internetinis paieškos robotas buvo paleistas 57 kartus. Savo darbo metu jis aplankė 200 skirtingų svetainių. Siekiant nustatyti svetainių svarbą kuo greičiau, tyrimo metu naudotas pirminės svetainių analizės paieškos robotas. Apie jį plačiau galite skaityti šio dokumento 3.9.3 skyriuje „Paieškos proceso struktūra“.

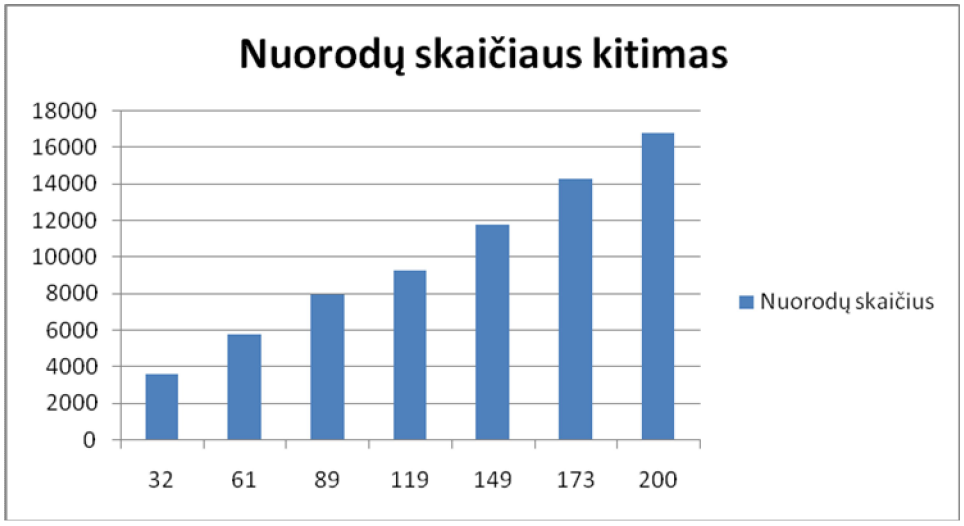
Žemiau pateikiamos paieškos proceso metu sukauptos informacijos (svetainių, nuorodų, svetainių tarpusavio sąryšio) pokyčių diagramos. Diagramų abscisių ašyje pateikiamas paieškos proceso aplankytų svetainių skaičius, ordinačių ašyje – sukauptos informacijos skaičius.

Iš diagramų galime matyti, kad vienos svetainės aplankymo metu paieškos robotas užregistruoja vidutiniškai 6-is naujų svetainių adresus, 84-ias svetainių vidines nuorodas, 9-is svetainių tarpusavio sąryšius.

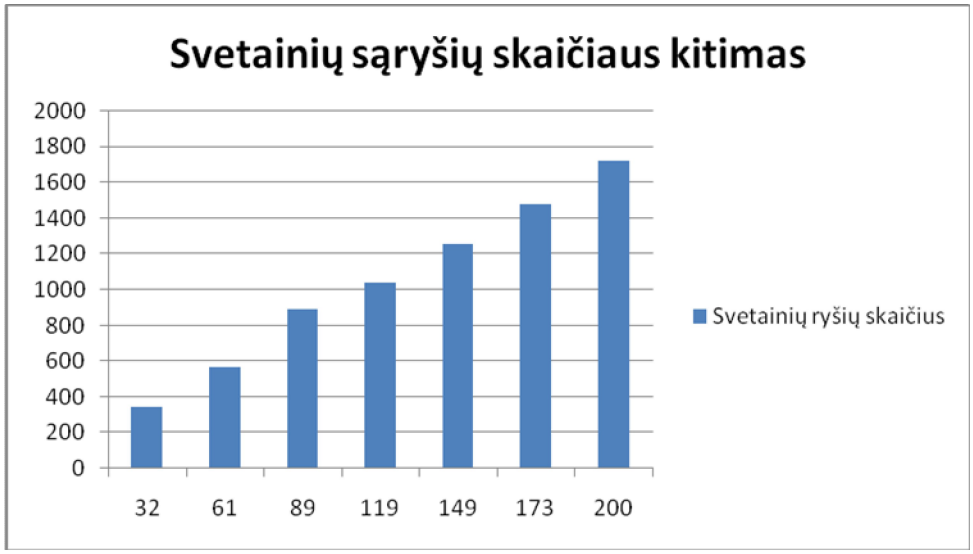
Internetinio paieškos proceso tyrimo metu surinkta 2011-05-19 dienos svetainių informacija.



5.1 pav. Svetainių skaičiaus augimas paieškos proceso metu



5.2 pav. Svetainių vidinių nuorodų skaičiaus augimas paieškos proceso metu



5.3 pav. Svetainių tarpusavio sąryšių skaičiaus kitimas paieškos proceso metu

Žemiau pateikta tyrimui pasirinktų svetainių išmatuotų metrikų reikšmių lentelė. Metrikų reikšmės nustatytos po internetinio paieškos roboto 200 skirtingų svetainių aplankymo.

Svetainių NNS reikšmė gauta po pakartotinio svetainių aplankymo praėjus 5 valandoms.

Svetainių DG reikšmė priklauso nuo domeno galūnės, žiūrėti 4.4 lentelę.

5.1 lentelė. Išmatuotos nagrinėjamų svetainių metrikų reikšmės

Nr.	Svetainė	ĮSNS	ISNS	VNSPS	NNS	DG
1.	http://www.delfi.lt	3	39	206	46	3
2.	http://www.skelbiu.lt	3	0	152	0	3
3.	http://www.alfa.lt	2	38	459	92	3
4.	http://www.pazintys.lt	9	3	141	34	3
5.	http://www.krepsinis.net/	2	18	208	12	2
6.	http://www.autogidas.lt	2	2	89	0	3
7.	http://www.musulaikas.com	5	32	46	0	3
8.	http://www.microsoft.com	0	28	194	0	3
9.	http://pigu.lt	0	2	398	14	3
10.	http://www.ktu.lt/	0	5	32	0	3

Tyrime naudojamas svetainių grupavimas pagal svetainės adreso domeno tipą. Tai leido suteikti tam tikroms labiau dominančioms svetainėms (šiuo atveju adresams „com“, „eu“ ir „lt“) didesnę svarbą. Domenų grupių svorinius koeficientus galite rasti 4.4 lentelėje.

5.4 lentelė. Domenų galūnių svoriniai koeficientai

Domeno galūnė	Svorinis koeficientas
com, eu, lt	3
org, net, biz	2
Visos kitos galūnės	1

5.2. Eksperimento eiga

Eksperimento eigos metu buvo atlikti trys bandymai su skirtingomis svorinių koeficientų kombinacijomis. Kiekvieno bandymo metu akcentuojamos tam tikros rūšies metrikos (pvz. svetainių turinio metrikos). Po kiekvieno bandymo pateikiama paieškos roboto nustatytų svetainių svarbos reikšmių lentelė.

Pirmas bandymas

Pirmajam bandymui buvo pasirinktos tokios svorinių koeficientų reikšmės, kurios akcentuoja svetainių tarpusavio ryšių ĮSNS ir ISNS metrikas. Pasirinktas svorinių koeficientų reikšmes galite matyti 4.5 lentelėje. Įeinančių ir išeinančių nuorodų svorinė vertė yra lygiavertė. Taip pat lygiavertės ir vidinių nuorodų bei naujų užregistruotų nuorodų svorinės reikšmės.

5.5 lentelė. Svetainių svarbos metrikų svorinių koeficientų reikšmės (1 bandymas)

Metrikos pavadinimas	Svorinis koeficientas
Įeinančių svetainių nuorodų skaičius (ĮSNS)	100
Išeinančių svetainių nuorodų skaičius (ISNS)	100
Vidinių nuorodų skaičius pradiniame puslapyje (VNSPS)	1
Domeno tipas (DT)	1000
Naujų nuorodų skaičius (NNS)	1

Žemiau esančiame 4.4 paveiksle pateikiamos paieškos proceso nustatytos tyrinėjamų svetainių svarbos reikšmės.

Nr.	Svetainė	Svarbos metrikų taškų suma	Svarbos reikšmė
1.	http://www.alfa.lt	7851	8
2.	http://www.delfi.lt	7452	7
3.	http://www.musulaikas.com	6746	7
4.	http://www.microsoft.com	5994	6
5.	http://www.pazintys.lt	4375	4
6.	http://www.krepsinis.net	4220	4
7.	http://pigu.lt	3612	4
8.	http://www.ktu.lt	3532	3
9.	http://www.autogidas.lt	3489	3
10.	http://www.skelbiu.lt	3452	3

5.4 pav. Išmatuotos svetainių svarbos reikšmės (1 bandymas)

Pastaba: svetainės svarbos reikšmė pateikiama dešimtbalėje sistemoje. Svarbos reikšmė nustatoma padalinus visą svarbos reikšmių imties intervalą į lygias dalis. Bendras svetainės svarbos taškų skaičius atitinka atskirų metrikų taškų sumą.

Kaip galime matyti iš 4.4 paveikslėlyje pateiktų rezultatų, didžiausią nustatytą svarbos reikšmę kaip ir galima buvo tikėtis turi svetainės turinčios daugiausiai įeinančių bei išeinančių svetainių nuorodų. Pirmųjų keturių svetainių atžvilgiu tai yra gerai, tačiau kaip matome svetainės http://www.skelbiu.lt bei http://pigu.lt, turinčios daug vidinių nuorodų yra labai nuvertinamos. Tai gali stipriai užvėlinti šių svetainių detalesnę vidinių nuorodų analizę bei pakartotinį aplankymą. Galima daryti išvadą, kad reikėtų sumažinti svetainių tarpusavio ryšių svorinius koeficientus arba padidinti nuorodų koeficientus.

Antras bandymas

Antrajam bandymui buvo pasirinkta kitokia metrikų svorinių koeficientų kombinacija. Šiuo atveju stipriai sumažinta įeinančių bei išeinančių svetainių svorinė vertė. Vis dėlto išeinančios svetainės laikomos svarbesnėmis. Dešimt kartų padidintos VNSPS ir NNS svorinių koeficientų reikšmės. Toks koeficientų pasiskirstymas leidžia suteikti didesnę svarbą svetainėms, turinčioms daug vidinių nuorodų bei dažnai atnaujinančioms turinį.

5.6 lentelė. Svetainių svarbos metrikų svorinių koeficientų reikšmės (2 bandymas)

Metrikos pavadinimas	Svorinis koeficientas
Įeinančių svetainių nuorodų skaičius (ĮSNS)	1
Išeinančių svetainių nuorodų skaičius (ISNS)	5
Vidinių nuorodų skaičius pradiniame puslapyje (VNSPS)	10
Domeno tipas (DT)	1000
Naujų nuorodų skaičius (NNS)	10

Nustatytas svetainių svorinių koeficientų reikšmes galite matyti 4.5 paveiksle.

Nr.	Svetainė	Svarbos metrikų taškų suma	Svarbos reikšmė
1.	http://www.alfa.lt	11702	9
2.	http://pigu.lt	7130	6
3.	http://www.delfi.lt	5718	5
4.	http://www.microsoft.com	5080	4
5.	http://www.pazintys.lt	4774	4
6.	http://www.skelbiu.lt	4523	4
7.	http://www.krepsinis.net	4292	3
8.	http://www.autogidas.lt	3902	3
9.	http://www.musulaikas.com	3625	3
10.	http://www.ktu.lt	3345	3

5.5 pav. Išmatuotos svetainių svarbos reikšmės (2 bandymas)

Kaip galime matyti iš antrojo bandymo rezultatų, vėlgi didžiausią svarbą išlaikė didieji informaciniai portalai http://www.alfa.lt/ bei http://www.delfi.lt. Į antrąją vietą dėl didelio vidinių nuorodų skaičiaus pateko svetainė http://pigu.lt. Nustatyti svetainių svarbos koeficientai yra pakankamai geri, tačiau reikėtų atkreipti dėmesį devintoje vietoje atsidūrusią svetainę http://www.musulaikas.com, kuri turi 37 svetainių tarpusavio nuorodas. Iš to galima daryti išvadą, kad reikėtų šiek tiek padidinti ISNS bei ĮSNS metrikų svorinių koeficientų reikšmes.

Trečiasis bandymas

Trečiajam bandymui buvo pasirinktas svorinių koeficientų rinkinys, įvertinantis tiek svetainių tarpusavio ryšio metrikas tiek ir svetainių turinio metrikas. Metrikų svorinių koeficientų reikšmės pateiktos 4.7 lentelėje.

5.7 lentelė. Svetainių svarbos metrikų svorinių koeficientų reikšmės (3 bandymas)

Metrikos pavadinimas	Svorinis koeficientas
Įeinančių svetainių nuorodų skaičius (ĮSNS)	10
Išeinančių svetainių nuorodų skaičius (ISNS)	100
Vidinių nuorodų skaičius pradiniame puslapyje (VNSPS)	5
Domeno tipas (DT)	1000
Naujų nuorodų skaičius (NNS)	100

Nr.	Svetainė	Svarbos metrikų taškų suma	Svarbos reikšmė
1.	http://www.alfa.lt	11525	10
2.	http://www.delfi.lt	8405	7
3.	http://www.microsoft.com	6770	6
4.	http://www.musulaikas.com	6455	6
5.	http://pigu.lt	5330	5
6.	http://www.krepsinis.net	4970	4
7.	http://www.pazintys.lt	4390	4
8.	http://www.skelbiu.lt	3775	3
9.	http://www.ktu.lt	3660	3
10.	http://www.autogidas.lt	3655	3

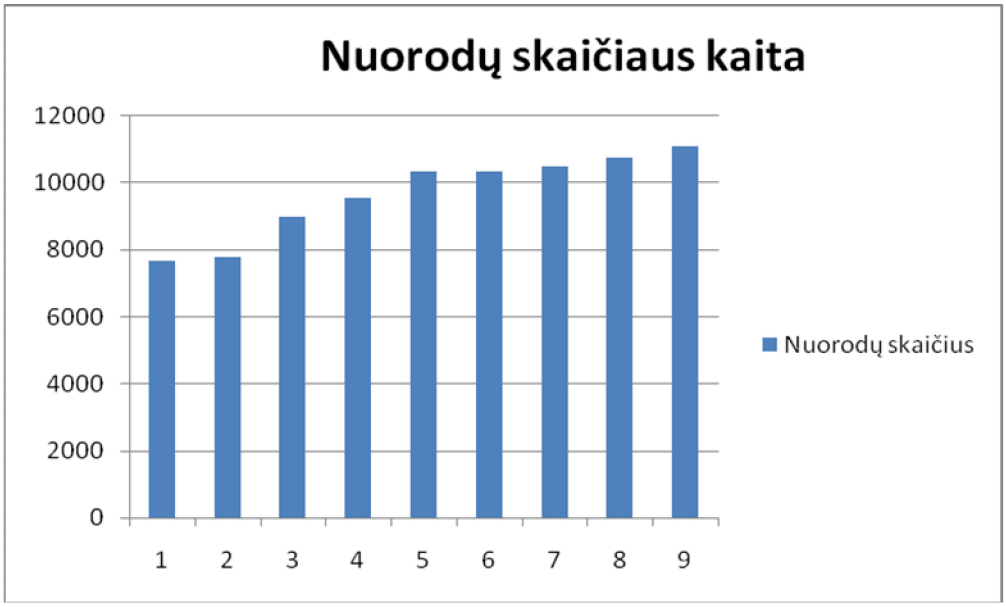
5.6 pav. Išmatuotos svetainių svarbos reikšmės (3 bandymas)

Iš gautų rezultatų galite matyti, kad šiuo atveju svetainės išdėstė tolygiau abiejų tipų metrikų atžvilgiu. Ši nustatytų svorinių koeficientų kombinacija buvo panaudota tolimesniame tyrime.

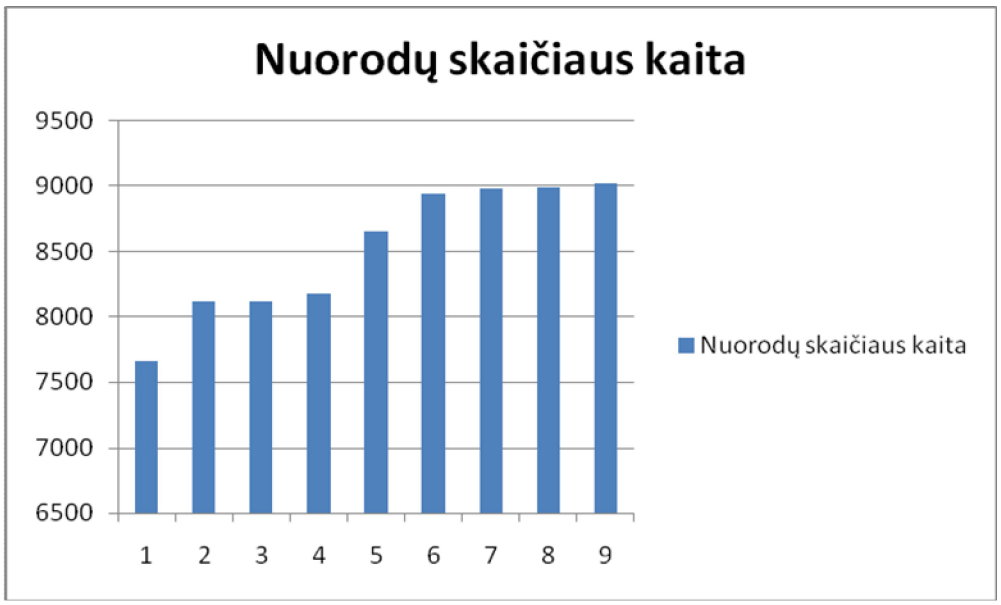
Norėdamas patikrinti kaip stipriai svetainių svarbos nustatymo algoritmo naudojimas įtakoja paieškos proceso efektyvumą atlikau svetainių analizės bandomąjį testą. Abiem atvejais buvo naudotas tas pats pradinių svetainių rinkinys. Pirmu atveju geresnėms svetainėms išskirti naudotas svetainių svarbos nustatymo algoritmas (žiūrėti 4.7 pav.). Antru atveju naudotas standartinis paieškos procesas be svetainių svarbos vertinimo (žiūrėti 4.8 pav.).

Abiem atvejais bandymo metu naudotas vidinės svetainių analizės paieškos robotas. Buvo nustatyta, kad jis analizuotų po 25 kiekvienos svetainės vidines nuorodas. Tada jis pereina prie kitos nagrinėjamos svetainės. Viso buvo išnagrinėta du šimtai 8 skirtingų svetainių vidinių nuorodų.

Kaip matome iš rezultatų, analizuodamas didesnio svarbumo svetaines, paieškos robotas užregistravo akivaizdžiai daugiau nuorodų nei eidamas paeiliui per visą svetainių sąrašą. Užregistruotų nuorodų kiekio skirtumas po to paties žingsnių skaičiaus skiriasi beveik 2000. Galima daryti išvadą, kad svetainių svarbos koeficientų nustatymo algoritmas stipriai pagerino internetinio paieškos proceso efektyvumą.



5.7 pav. Užregistruotų nuorodų kaita (naudojamas svetainių svarbos nustatymo algoritmas)



5.8 pav. Užregistruotų nuorodų kaita (paieškos procesas be svetainių svarbos nustatymo)

5.3. Rezultatai

Magistrinio darbo tyrimo bei eksperimento metu buvo atlikta svetainių svarbos svorinių koeficientų analizė. Buvo atlikti trys bandymai su skirtingomis svorinių koeficientų reikšmių kombinacijomis. Po kiekvieno bandymo buvo pateikiama tiriamų svetainių nustatytų svarbos reikšmių lentelė.

Kaip galime matyti iš gautų rezultatų, nustatytos svetainių svarbos reikšmės daugiau ar mažiau skyrėsi nuo google PR rezultatų (žiūrėti 4.1 lentelę). Taip yra todėl, kad kai kurių svarbos metrikų nustatymui, pvz. NNS ir ĮSNS, reikia įvykdyti daugiau žingsnių, neužtenka atlikti vien tik pirmo svetainės puslapio analizės. Kitas rezultatų skirtumą įtakojantis veiksnys yra tai, kad mano siūlomas svetainių svarbos nustatymo algoritmas skirtas nustatyti svetainių svarbą po pirmo arba keleto svetainės aplankymų. Akivaizdu, kad tiriamą svetainę aplankius daugiau kartų šie rezultatai gali keistis.

Svetainių svarbos skaičiavimui buvo naudojami atskirų metrikų svoriniai koeficientai. Svarbos reikšmė buvo nustatoma susumavus visų svorinių koeficientų ir išmatuotų metrikų sandaugų porų sumą.

Buvo nustatyta, kad optimaliausias svorinių koeficientų rinkinys iš visų trijų bandymų, yra paskutinis. Šiuo atveju svoriniai koeficientai buvo pasirinkti remiantis prielaida, kad didesnės svarbos svetainės turi daugiau išeinančių kitų svetainių nuorodų bei dažniau atnaujina savo turinį. Žinoma, esant poreikiui, galima naudoti ir kitokias svorinių koeficientų kombinacijas.

Buvo atliktas tyrimas kaip svorinių koeficientų naudojimas įtakoja paieškos proceso spartą. Gauti rezultatai 4.7 ir 4.8 pav., kaip ir buvo galima tikėtis, akivaizdžiai parodo svetainių svarbos skaičiavimo naudojimo pranašumą.

6. Išvados

- Magistrinio darbo metu realizuotas internetinis paieškos procesas, atliekantis pagrindines tokio tipo sistemų funkcijas. Sukurta sistema leis atlikti internetinių svetainių struktūros bei turinio analizę, suteiks galimybę vykdyti vartotojo pasirinktų dokumentų paiešką bei svetainių turinio atsisiuntimą.
- Literatūros analizės metu aprašyti pagrindiniai internetinio paieškos roboto architektūros elementai, nustatyta, kad didžiausios paieškos proceso problemos kyla dėl milžiniško interneto dydžio bei nuolatinės informacijos kaitos.
- Programinė sistema realizuota viena populiariausių internetinės srities programavimo kalbų – PHP. Sistemos struktūra remiasi MVC architektūros principais, kas leido atskirti informacijos atvaizdavimo sluoksnį nuo informacijos saugojimo bei apdorojimo.
- Sistemos esminiai komponentai yra šie: sistemos branduolys, vartotojo sąsajos generavimo posistemė ir duomenų apdorojimo posistemė.
- Realizuotas internetinis paieškos procesas yra skaidomas į 4 mažesnius komponentus: pirminės svetainių analizės vorą, vidinių nuorodų analizės vorą, pakartotinio lankymo vorą ir svetainių svarbos nustatymo procesą. Tai leido labiau sukonkretinti atskirų paieškos proceso komponentų atliekamas funkcijas.
- Magistrinio darbo tyrimo metu buvo atlikta svetainių svarbos skaičiavimo metrikų analizė. Buvo nustatyta, kad skaičiuojant svetainių svarbos reikšmę labiausiai reikia atsižvelgti į išeinančių svetainių nuorodų skaičių, vidinių nagrinėjamos svetainės nuorodų skaičių bei turinio atnaujinimo dažnumą.
- Svetainių svarbos koeficientų naudojimas leido stipriai padidinti paieškos proceso efektyvumą.

7. Literatūra

[1] Matthew Gray. Internet Growth and Statistics: Credits and Background. [Žiūrėta 2011 05 16], prieiga internete: http://www.mit.edu/people/mkgray/net/background.html

[2] The first web robot – 1993. [Žiūrėta 2011 05 16], prieiga internete: http://www.salientmarketing.com/seo-resources/search-engine-history/web-robot.html

[3] „Web Crawlers", Jury Byrd. [Žiūrėta 2009 11 05], prieiga internete: http://cseweb.ucsd.edu/~paturi/cse91/Presents/jbyrd.pdf

[4] „Implementing an effective web crawler". [Žiūrėta 2011 05 16], prieiga internete: http://www.ennovatetech.com/downloads/webcrawler.pdf

[5] Bots vs Browsers - database of 672,480 user agents. [Žiūrėta 2011 05 16], prieiga internete http://www.botsvsbrowsers.com/

[6] „Web crawlers. Googlebot". [Žiūrėta 2011 05 16], prieiga internete http://www.milkaddict.com/web-crawlers-googlebot/

[7] „How Google Works". [Žiūrėta 2009 11 05], prieiga internete http://www.googleguide.com/google_works.html

[8] „Optimizing for Bing in 2010". [Žiūrėta 2011 05 16], prieiga internete: http://www.marketing-jive.com/2010/09/optimizing-for-bing-in-2010-part-iii.html

[9] "The Yahoo SLURP Crawler", Akinola Akintomide. [Žiūrėta 2009 11 05], prieiga internete: http://www.seochat.com/c/a/Search-Engine-Spiders-Help/The-Yahoo-SLURP-Crawler/

[10] „Web crawler". [Žiūrėta 2011 05 16], prieiga internete: http://www.interpriseo.com/resources/general_info_articles/Web%20Crawler.pdf

[11] „UbiCrawler: A Scalable Fully Distributed Web Crawler", Paolo Boldi, Bruno Codenotti, Massimo Santini, Sebastiano Vigna. [Žiūrėta 2011 05 16].
Prieiga internete http://vigna.dsi.unimi.it/ftp/papers/UbiCrawler.pdf

[12] „Design and Implementation of a High-Performance Distributed Web Crawler", Vladislav Shkapenyuk ir Torsten Suel. [Žiūrėta 2009 11 06],
prieiga internete: http://cis.poly.edu/suel/papers/crawl.pdf

[13] „Measuring The Web Crawler Ethics“, C. Lee Giles, Yang Sun ir Isaac G. Councill. [Žiūrėta 2011 05 16].

[14] „Implementation of Focused Crawler“, Yang Yongsheng, Wang Hui. [Žiūrėta 2009 11 09], prieiga internete: http://www.cs.ust.hk/~ysyang/courses/comp630d/630dreport.pdf

[15] „Web crawler“. [Žiūrėta 2011 05 16].

[16] „Do Your Worst to Make the Best: Paradoxical Effects in PageRank Incremental Computations”, Paolo Boldi, Massimo Santini, Sebastiano Vigna. [Žiūrėta 2011 05 16].

[17] „Web Crawler On Client Machine“, Rajashree Shettar, Dr. Shobha G. [Žiūrėta 2009 11 13], prieiga internete: http://www.iaeng.org/publication/IMECS2008/IMECS2008_pp1121-1124.pdf

[18] „Web Crawler Architecture”, Marc Nayork. [Žiūrėta 2009 11 13], prieiga internete: http://research.microsoft.com/pubs/102936/EDS-WebCrawlerArchitecture.pdf

[19] A New Crawling Model and Architecture. [Žiūrėta 2011 05 16], prieiga internete: http://chato.cl/papers/crawling_thesis/newmodel.pdf

[20] „Balancing Volume, Quality and Freshness in Web Crawling“, Ricardo Baeza-Yates and Carlos Castillo. [Žiūrėta 2009 11 14], prieiga internete: http://www.ciw.cl/recursos/webCrawling.pdf

[21] „Practical Web Crawling Issues“, Carlos Castillo. [Žiūrėta 2009 11 14], prieiga internete: http://www.chato.cl/papers/crawling_thesis/practical.pdf

[22] „Web Crawler“, Joseph Jang. [Žiūrėta 2009 11 15], prieiga internete: http://hwanjoyu.org/teaching/2009-1-introdm/project/medline/web_crawler_20070223.pdf

8. Terminų ir santrumpų žodynas

IT – informacinės technologijos.

HTML – tai kompiuterinė žymėjimo kalba, naudojama pateikti turinį internete.

DNS – sistema, kuri taikoma susieti duomenis su internetiniu adresu. Taip pat ši sistema transformuoja informacinius skaitinius adresus į žmonėms priimtinesnį pavidalą.

IP – kompiuterio identifikatorius.

MVC (Model-View-Controller) – tai sistemos architektūros modelis, padalinantis programinį kodą į keletą lygių.

PHP – plačiai paplitusi dinaminė interpretuojama programavimo kalba dažniausiai naudojama internetinėms svetainėms kurti.

MySQL – reliacinė duomenų bazių valdymo sistema, palaikanti daugelį naudotojų, dirbanti SQL kalbos pagrindu.

RAM – kompiuterio (serverio) operatyvioji atmintis.

HDD – kompiuterio kietasis diskas (duomenų saugykla).

PR (pagerank) – kriterijus, apibūdinantis svetainės svarbą.

Xampp – internetinio Apache serverio versija, kurią galima įdiegti asmeniniame kompiuteryje.

Wamp – internetinio serverio programinė įranga, kurią galima naudoti asmeniniame kompiuteryje.

Javascript – objektiškai orientuota skriptų programavimo kalba, besiremianti prototipų principu.

Ajax – terminas, apibrėžiantis svetainių programavimo technologiją, naudojančią šias priemones maksimaliam interaktyvumui pasiekti.

Unix – operacinių sistemų grupė.

Flash – daugialypė programinė įranga, sukurta Macromedia kompanijos.

iFrame – HTML kalbos objektas, leidžiantis svetainėje integruoti kitos svetainės langą.

HTTP – pagrindinis protokolas naudojamas pasiekti informaciją pasauliniame tinkle (WWW).

Hidden (deep) web – interneto sritis, dažniausiai tiesiogiai neprieinama internetinių paieškos robotų.

Cronjob – periodinės užklausos, vykdomos internetiniame serveryje.

Volere šablonai – standartizuoti šablonai, naudojami tam tikros srities informacijai aprašyti, pavyzdžiui: funkciniams reikalavimams.

9. Priedai

9.1. Licencija

Ši licencija apibrėžia magistrinio projekto – internetinio paieškos proceso – naudojimo taisykles bei asmenų, turinčių sistemos licenciją, teises ir įsipareigojimus. Magistrinis projektas tolimesniame tekste vadinamas „Programa“. Į kiekvieną licencijos turėtoją tekste kreipiamasi „Jūs“.

Šią Programą Jūs galite naudoti tik nekomerciniais tikslais. Bet kokie bandymai siekti komercinės naudos platinant Programą pažeidžia šios licencijos sąlygas bei panaikina Jūsų teises į šį produktą.

Draudžiama naudoti Programą ir jos sukauptus duomenis neteisėtiems tikslams bei veikloms, kurios draudžiamos Lietuvos Respublikos bei tarptautinių įstatymų. Draudžiama naudoti Programą siekiant trikdyti kitų sistemų darbą arba kitaip neigiamai paveikti kitų organizacijų veiklą.

Jūs galite kopijuoti ir platinti originalius Programos išeities tekstus bet kokiose laikmenose, kuriose Jūs juos gavote ar patys patalpinote, aiškiai ir kaip priklauso kiekvienoje kopijoje įtraukdami atitinkamus autorinių teisių įspėjimus. Nekeiskite jokių įspėjimų susijusių su šia licencija ir visiems Programos gavėjams pateikite šios licencijos originalo kopiją kartu su Programa.

Jūs galite laisvai modifikuoti Programą arba atskiras jos dalis siekdami pritaikyti Programą savo poreikiams.

Jūs negalite kopijuoti, licencijuoti ar platinti Programos kitaip nei aiškiai numatyta šioje licencijoje. Bet kokie bandymai kitaip kopijuoti, licencijuoti ar platinti Programą yra negaliojantys ir automatiškai panaikina Jūsų teises suteiktas šios licencijos. Kitų asmenų, gavusių iš Jūsų kopijas ar teises remiantis šia licencija, teisės (licencijos) nebus panaikintos, jei šie asmenys nepažeidė licencijos.

Kiekvieną kartą, kai Jūs platinate Programą (ar bet kokį Programa paremtą darbą), Programos gavėjas automatiškai gauna licenciją iš pirmojo Programos autoriaus, suteikiančią teisę kopijuoti, platinti ar modifikuoti Programą remiantis šiomis sąlygomis. Jūs negalite gavėjui primesti jokių papildomų apribojimų nesančių šioje licencijoje.