VILNIAUS UNIVERSITETAS
MATEMATIKOS IR INFORMATIKOS FAKULTETAS
PROGRAMŲ SISTEMŲ STUDIJŲ PROGRAMA

Paskirstyta saityno paieškos roboto sistema didelio masto modernių taikomųjų žiniatinklio programų rodyklių sudarymui

Distributed Web Crawling System for Large-Scale Indexing of Modern Web Applications

Bakalaurinio darbo planas

Atliko: 4 kurso 5 grupės studentas
Kasparas Taminskas (parašas)

Darbo vadovas: Lekt. Aurimas Šimkus (parašas)

Vilnius – 2020

TURINYS

1. TYRIMO OBJEKTAS IR AKTUALUMAS 3
2. DARBO TIKSLAS, UŽDAVINIAI, LAUKIAMI REZULTATAI 4
3. TYRIMO METODAS IR DARBO ATLIKIMO PROCESAS 5

1. Tyrimo objektas ir aktualumas

Saityno paieškos robotas – tai sistema, kurios tikslas dideliais mastais periodiškai ir sistematiškai lankyti Interneto svetaines ir masiškai parsisiųsti jų informaciją tam, kad būtų galima sukurti lokalų rodyklių[1] sąrašą [ON10]. Tokio tipo sistemos dažniausiai pritaikomos renkant duomenis Interneto paieškos sistemų variklių algoritmams, kurie vartotojams pasiūlo aktualius paieškos rezultatus pagal pateiktas paieškos užklausas (pvz.: „Google“, „Bing“, „Yahoo!“, „Baidu“ komercinės paieškos sistemos). Paieškos robotų pritaikymo galimybės itin didelės – jie taip pat naudojami sudarant Interneto svetainių archyvus arba atliekant duomenų gavybą[2] statistiniams tyrimams, mašininio mokymosi duomenų rinkinių kaupimui. Šios sistemos gali būti naudojamos kartu su Saityno informacijos surinkimo technologijomis[3] rinkodaros, įdarbinimo, verslo konkurentų žvalgybos sferose, nelegalaus, plagijuoto turinio paieškai.

Paieškos robotų technologijos tema nėra nauja programų sistemų inžinerijos industrijoje – tokios programos atsirado su pirmosiomis Interneto naršyklėmis dar XX a. 10-ame dešimtmetyje siekiant sudaryti efektyvius paieškos indeksus ir buvo nuolat tobulinamos. Iš pirmo žvilgsio sprendžiama problema gana paprasta – paieška į plotį[4] paremta svetainių lankymo sistema [ON10]. Paprastumą komplikuoja faktas, jog Interneto svetainių, jas sudarančių puslapių skaičius nenustoja didėti eksponentiniais greičiais (palyginimui – 2016 metais buvo apie 1 mlrd. Interneto svetainių, o 2017 metais – jau 1,76 mlrd.) [STR+20]. Tokio tempo augimas reikalauja, kad suprojektuotos paieškos robotų sistemos sugebėtų efektyviai rinkti ir saugoti didžiulius kiekius duomenų, būtų lengvai plečiamos – horizontalus resursų prijungimas suteiktų tiesinės priklausomybės efektą). Riboti duomenų saugojimo talpyklų resursai verčia galvoti ir apie duomenų aktualumo, atsikratymo sistematikas [ON10]. Nors teorinių mokslinių straipsnių apie paieškos robotų architektūras netrūksta, praktinių efektyvios, paskirstytos sistemos kūrimo ir tikslaus svetainių aplankymo efektyvumo matavimo tyrimų pasigendama. Padėtį stipriai apsunkina naujos kartos modernios taikomosios interneto programos, kurių negalima vadinti paprastomis interneto svetainėmis – jas talpinantys serveriai negrąžina viso turinio su HTTP protokolo užklausa, o turinys inkrementiškai ir dinamiškai įterpiamas aplikacijos vykdymo metu.[5] [DFK+08] Tokio turinio nuskaitymas tampa kompleksiška užduotimi. Komercinės paieškos sistemos linkusios slėpti arba pateikti labai abstrakčias, dalines savo paieškos robotų architektūras, jų implementacijos technologijas, nes jos nemaža dalimi lemia šių kompanijų strateginę sėkmę.

[1] angl. – Web Index
[2] angl. – Data Mining
[3] angl. – Web Scraping - struktūrizuotas puslapio duomenų surinkimas
[4] angl – Breadth-First-Search
[5] angl. – Single Page Applications - turinį dinamiškai užkrauna vykdomas Javascript kodas

2. Darbo tikslas, uždaviniai, laukiami rezultatai

Darbo tikslas – Ištirti ir pasiūlyti galimą pasaulinio modernių taikomųjų saityno programų tinklo išskirstyto saityno paieškos roboto sprendimą panaudojant debesų kompiuterijos galimybes

Uždaviniai

Siekiant išsikelto rašto darbo tikslo, išsikelti šie pagrindiniai darbo uždaviniai:

1. Identifikuoti ir apibrėžti Saityno paieškos roboto sistemų (angl. – Web Crawling Systems) atsakomybių ribas, akcentuoti pagrindinius skirtumus nuo Saityno informacijos surinkimo sistemų (angl. – Web Scraping Systems)
2. Išnagrinėti šiuolaikinių taikomųjų saityno programų žvalgymo naudojant saityno paieškos robotus pagrindinius keliamus iššūkius
3. Išanalizuoti kelias viešai prieinamas paieškos robotų sistemų architektūras, akcentuoti, kaip jos siūlo spręsti apibrėžtus iššūkius, išskirti bendras tokių sistemų komponentes
4. Atlikti viešų debesų kompiuterijos paslaugų tiekėjų siūlomų servisų paiešką, išrinkti tinkamų paslaugų poaibį paieškos roboto sistemos iššūkių sprendimui
5. Apibrėžti debesų kompiuterijos technologijomis paremto prototipo architektūrinį dizainą
6. Realizuoti apibrėžtą žiniatinklio paieškos roboto sistemos prototipą
7. Įvertinti prototipinės realizacijos panaudojamumo, plečiamumo, pritaikymo pagal specifinius poreikius galimybes

Laukiami rezultatai

1. Apibrėžtos žiniatinklio paieškos roboto atsakomybių sritys, kuriomis bus apsiribojama prototipo realizacijoje, išskirti esminiai skirtumai nuo Saityno informacijos surinkimo sistemų
2. Išskirti pagrindiniai modernių taikomųjų saityno programų žvalgymo iššūkiai
3. Identifikuotos ir aprašytos bendros ir daugumoje šaltinių vieningai minimos žiniatinklio paieškos roboto sistemos architektūros komponentės
4. Išrinktas poaibis debesų kompiuterijos paslaugų tiekėjo siūlomų paslaugų, kuriuo remiantis bus projektuojamas ir realizuojamas prototipas
5. Grafiškai (UML diagramos) apibrėžtas projektuojamos žiniatinklio paieškos roboto sistemos dizainas, pasirinktos realizacijos technologijos – programavimo kalba, karkasai
6. Pagal apibrėžtą dizainą realizuotas ir kodo versijavimo sistemoje patalpintas viešai prieinamas prototipas
7. Sukurtas prototipas išbandytas praktiškai koreguojant sistemos resursų galimybes, grafiškai pateikti eksperimento rezultatai

3. Tyrimo metodas ir darbo atlikimo procesas

Tyrimo metodas

Rašto darbe bus renkama ir analizuojama literatūrinė medžiaga, susijusi su žiniatinklio paieškos robotų sistemomis, jų architektūra, galimomis realizacijos technologijomis. Taip pat bus vykdoma alternatyvių žiniatinklio paieškos roboto sistemų architektūrų palyginamoji analizė. Praktinėje dalyje bus atliekamas eksperimentinis tyrimas, kurio pagrindinis tikslas – įvertinti sukurto prototipo panaudojamumo, plečiamumo galimybes.

Numatomas darbo atlikimo procesas

1. Teorinė žiniatinklio paieškos robotų apžvalga: atsiradimo priežastys, atliekamos funkcijos, skirtumai nuo žiniatinklio informacijos surinkimo sistemų, pagrindiniai iššūkiai, su kuriais susiduriama šiomis dienomis vykdant žiniatinklio žvalgybą tokiomis sistemomis
2. Analizuojamos ir lyginamos skirtingos viešai prieinamos žiniatinklio paieškos robotų sistemų architektūros [HN99] [BCS+04] – identifikuojami bendriniai tokių sistemų komponentai
3. Atliekama galimų debesų kompiuterijos paslaugų tiekėjų siūlomų servisų, platformų analizė, kurios padėtų įgyvendinti esminius prototipinės sistemos architektūros komponentus [Cor20] [Ser20]
4. Išskyrus bendrinius tokių sistemų komponentus ir apsibrėžus aibę naudotinų debesų kompiuterijos paslaugų tiekėjo servisų sudaroma prototipo architektūra – braižomos UML struktūrinės ir elgsenos diagramos, taip pat kitokio formato diagramos, parodančios aukšto lygio sistemos vaizdą, komunikaciją tarp posistemių
5. Praktinė dalis – pagal apibrėžtą dizainą su pasirinktomis technologijomis, programavimo kalbomis, karkasų rinkiniu realizuojamas paskirstytos žiniatinklio paieškos roboto sistemos prototipas
6. Vykdomas realizuoto prototipo naudojimo eksperimentas – paduodamas pradinis interneto svetainių adresų rinkinys (startinis taškas), stebimas ir matuojamas aplankomų svetainių kiekis per laiko vienetą, sistemos dalių resursų apkrova (eilių perpildymas, procesoriaus, operatyviosios atminties, disko vietos naudojimas). Identifikuojamos kylančios galimų nesėkmių priežastys. Taip pat stebima, kaip keičiasi sistemos darbo efektyvumas horizontaliai didinant sistemos resursus – vertinamas prototipo plečiamumas
7. Atlikus eksperimentą ir apdorojus tyrimo rezultatus pateikiami grafikai ir lentelės, iš kurių daromos rašto darbo išvados apie sukurto prototipo panaudojimo perspektyvas, esminius privalumus ir trūkumus, nurodomos galimos ateities darbo kryptys

Naudotinos literatūros sąrašas

[BCS+04] Paolo Boldi, Bruno Codenotti, Massimo Santini ir Sebastiano Vigna. Ubicrawler: a scalable fully distributed web crawler. http://ccc.inaoep.mx/~villasen/bib/crawler.pdf, 2004. tikrinta 2020-03-09. Java kalba realizuotas dar vienas saityno paieškos robotas. Akcentas – ši architektūra pasižymi tuo, kad yra pilnai išskirstyta.

[Cor20] Microsoft Corporation. Azure services documentation. https://docs.microsoft.com/en-us/azure/, 2020. tikrinta 2020-03-09. Azure debesų kompiuterijos paslaugų tiekėjo dokumentacijos šaltinis padės susigaudyti ieškant nagrinėjamos problemos sprendimo technologijų pasiūlymų, juos realizuojant.

[DFK+08] Cristian Duda, Gianni Frey, Donald Kossmann ir Chong Zhou. Ajaxsearch: crawling, indexing and searching web 2.0 applications. https://web.archive.org/web/20120508001129/http://www.vldb.org/pvldb/1/1454195.pdf, 2008. tikrinta 2020-03-09. Pristatomi sprendimai, kaip būtų galima žvalgyti naujos kartos Saityno 2.0 puslapius, kurių turinys užkraunamas dinamiškai Javascript kodo pagalba.

[HN99] Allan Heydon ir Marc Najork. Mercator: a scalable, extensible web crawler. http://www.bagualu.net/linux/crawler.pdf, 1999. tikrinta 2020-03-09. Java kalba realizuotas plečiamas, papildomas saityno paieškos robotas – dar vienas architektūrinis pavyzdys, atskleidžiantis esminius tokių sistemų sudedamuosius komponentus.

[ON10] C. Olston and Marc Najork. Web crawling. http://infolab.stanford.edu/~olston/publications/crawling_survey.pdf, 2010. tikrinta 2020-03-09. Išsamus Stanfordo universiteto šaltinis, kuriame detaliai pristatoma paieškos robotų sistemų priešistorė, nagrinėjama tokių sistemų architektūra, pateikiami pagrindiniai iššūkiai, nurodomos tolesnio tyrimo kryptys. Šis šaltinis padeda geriau suvokti nagrinėjamos technologijos architektūrą.

[Ser20] Amazon Web Services. Amazon web services documentation. https://docs.aws.amazon.com/, 2020. tikrinta 2020-03-09. Amazon AWS debesų kompiuterijos paslaugų tiekėjo dokumentacijos šaltinis kaip alternatyva Azure paslaugoms.

[STR+20] F. M. Javed Mehedi Shamrat, Zarrin Tasni, A.K.M Sazzadur Rahman, Naimul Islam Nobel, and Syed Akhter Hossain. An effective implementation of web crawling technology to retrieve data from the world wide web (www). http://www.ijstr.org/final-print/jan2020/An-Effective-Implementation-Of-Web-Crawling-Technology-To-Retrieve-Data-From-The-World-Wide-Web-www.pdf, 2020. tikrinta 2020-03-09. Gana detali pavyzdinio saityno roboto architektūra ir jos pavyzdinė realizacija, kuri gali pagelbėti projektuojant išskirstytą sistemą.