DUOMENŲ ANALIZĖ IR VAIZDAVIMAS

Duomenų tyrybos sistemos, pagrįstos saityno paslaugomis

Olga Kurasova

Vilniaus universiteto Matematikos ir informatikos instituto daktarė
Vilnius University, Institute of Mathematics and Informatics, PhD, Senior researcher
Akademijos g. 4, LT-08663 Vilnius
El. paštas: Olga.Kurasova@mii.vu.lt

Virginijus Marcinkevičius

Vilniaus universiteto Matematikos ir informatikos instituto daktaras
Vilnius University, Institute of Mathematics and Informatics, PhD, Senior researcher
Akademijos g. 4, LT-08663 Vilnius
El. paštas: Virginijus.Marcinkevicius@mii.vu.lt

Viktor Medvedev

Vilniaus universiteto Matematikos ir informatikos instituto daktaras
Vilnius University, Institute of Mathematics and Informatics, PhD, Researcher
Akademijos g. 4, LT-08663 Vilnius
El. paštas: Viktor.Medvedev@mii.vu.lt

Aurimas Rapečka

Vilniaus universiteto Matematikos ir informatikos instituto doktorantas
Vilnius University, Institute of Mathematics and Informatics, PhD student
Akademijos g. 4, LT-08663 Vilnius
El. paštas: Aurimas.Rapecka@mii.vu.lt

Straipsnis skirtas duomenų tyrybos, pagrįstos saityno paslaugomis, analizei. Apibrėžiamos pagrindinės su saityno paslaugomis susijusios sąvokos. Pristatomos paskirstytosios duomenų tyrybos galimybės bei jų įgyvendinimo priemonės – Grid, Hadoop. Atliekama duomenų tyrybos sistemų, pagrįstų saityno paslaugomis, analitinė apžvalga. Parenkami sistemų palyginimo kriterijai. Pagal šiuos kriterijus atliekama populiariausių duomenų tyrybos sistemų, pagrįstų saityno paslaugomis, lyginamoji analizė. Nustatoma, kurios sistemos įvertinamos geriausiai, o kurios neatitinka daugumos kriterijų.

Įvadas

Duomenų tyryba (angl. *data mining*) yra svarbi žinių radimo duomenų bazėse proceso dalis. Kurį laiką tyrimai buvo nukreipti į duomenų tyrybos metodų kūrimą ir jų taikymą. Buvo intensyviai kuriamos įvairios duomenų tyrybos sistemos: *Weka* (Hall et al., 2009), *Orange* (Podpecan et al., 2012), *Knime* (Berthold et al., 2007), *RapidMiner* (Mierswa et al., 2006), R (Knell, 2013) ir kt. Pastaruoju metu sparčiai vystomi saityno paslaugų (angl. *web services*) tyrimai. Duomenų tyrybos sistemos kuriamos naudojant paslaugų architektūrą (angl. *Service-Oriented Architecture*, SOA). Be to, yra bandymų sukurti duomenų tyrybos algoritmus kaip saityno paslaugas, kurios gali būti panaudotos kitose sistemose, praplečiant jų galimybes be papildomo programavimo (FAEHIM, 2005). Dar viena aktuali duomenų tyrybos problema – didelių apimčių duomenų analizė, reikalaujanti ne tik specialių algoritmų, bet ir naujų technologijų.

Šio straipsnio tyrimo objektas – duomenų tyrybos sistemos, pagrįstos saityno paslaugomis, tyrimo metodika – analizuojamų sistemų lyginamoji analizė. Straipsnio tikslas – apžvelgti esamas duomenų tyrybos sistemas, pagrįstas saityno paslaugomis; parinkti kriterijus, pagal kuriuos atlikti šių sistemų lyginamąją analizę, ir išryškinti lyginamų sistemų pranašumus ir trūkumus. Tokia lyginamoji analizė bus naudinga kuriant naują duomenų tyrybos sistemą.

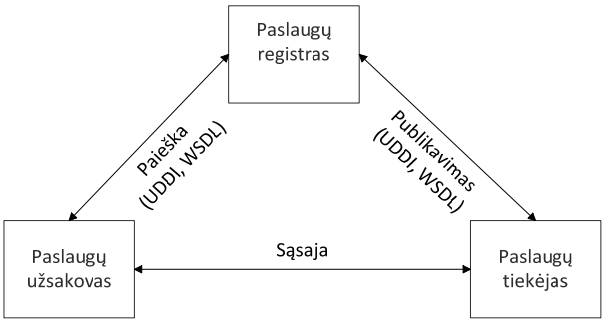
Paskirstytosios duomenų tyrybos įgyvendinimo priemonės

Pastaraisiais metais sparčiai plinta saityno paslaugos dėl jų paprasto panaudojimo ir plataus paslaugų stiliaus architektūros paradigmos taikymo įvairiose srityse. Paslaugų stiliaus architektūra – tai naujas programinės įrangos architektūros modelis, kuris užtikrina verslo procesų transformavimą į aibę susietų paslaugų (ar aibę pasikartojančių verslo užduočių), kurios gali būti pasiektos per saityną (IBM Developers Works, 2013). Transformavimas gali būti atliktas tiek lokaliame tinkle, tiek saityne ar serveryje. Saityno paslaugos jungiamos siekiant išspręsti specifines verslo užduotis ar vykdant sudėtingą verslo procesą. Verslo procesų sudėtingumą lemia tai, kad jiems vykdyti paprastai turi būti pasitelkta keletas skirtingų programinės įrangos paketų, kurie yra skirtingose kompiuterinėse platformose. SOA architektūra leidžia spręsti minėtas problemas naudojant programinės įrangos ir infrastruktūros perorganizavimo į aibę tarpusavyje bendraujančių paslaugų būdą (Papazoglou, 2003). Būtent tai užtikrina, jog saityno paslaugos gali būti greitai ir efektyviai pritaikytos prie kintančių verslo sąlygų.

Paslauga (angl. *service*) vadinamas programos blokas, gebantis atlikti specifines užduotis ar funkcijas, naudojant griežtai apibrėžtas sąsajas (IBM developers Works, 2013). Paslaugos veikia sistemos viduje, tačiau kiti sistemos komponentai nežino, kaip jos užtikrina savo funkcijas – svarbus tik paslaugos grąžinamas rezultatas. Saityno paslauga – tai viename komponente esantis funkcijų, pasiekiamų pasinaudojus standartiniais interneto protokolais, rinkinys. Paslaugos prieinamumas internetu suteikia galimybę integruoti skirtingose platformose veikiančias programas. Jos lengvai įdiegiamos dėl sukurtų saityno paslaugų standartų ir jų suderinamumo su visuotinai paplitusiomis ir naudojamomis technologijomis: XML (angl. *Extensible Markup Language*), SOAP (angl. *Simple Object Access Protocol*), REST (angl. *Representational State Transfer*), WSDL (angl. *Web Services Description Language*) ir UDDI (angl. *Universal Description Discovery and Integration*).

Kuriant saityno paslaugas ir naudojant paslaugų stiliaus architektūrą dalyvauja trys veikėjai: paslaugų registras, paslaugų tiekėjas ir paslaugų užsakovas (1 pav.). Paslaugų registre saugoma informacija apie siūlomas ar pasiekiamas saityno paslaugas, tam naudojant WSDL ir UDDI formatus. Paslaugų tiekėjai publikuoja savo siūlomas saityno paslaugas ir užtikrina jų funkcionalumą. Paslaugų užsakovas ieško tinkamų paslaugų, realizuojančių norimą funkcionalumą, ir integruoja su kitomis saityno paslaugomis, siekdamas įvykdyti tam tikrą procesą. Atliekant saityno paslaugos paiešką kreipiamasi į paslaugų registrą ir gaunamas bendras paslaugos aprašymas UDDI formatu arba surastos saityno paslaugos WSDL, kuriuo remiantis galima tiesiogiai kreiptis į paslaugos tiekėją.

Dažnai tiriami duomenys yra didelių apimčių. Nustatyta, kad 90 proc. visų pasaulio duomenų yra sukurta tik per pastaruosius dvejus metus. Todėl nuolat kyla poreikis analizuoti didelius duomenų kiekius. Tokių duomenų tyryba reikalauja



1 p a v. ***Saityno paslaugų architektūra (Kreger, 2001)***

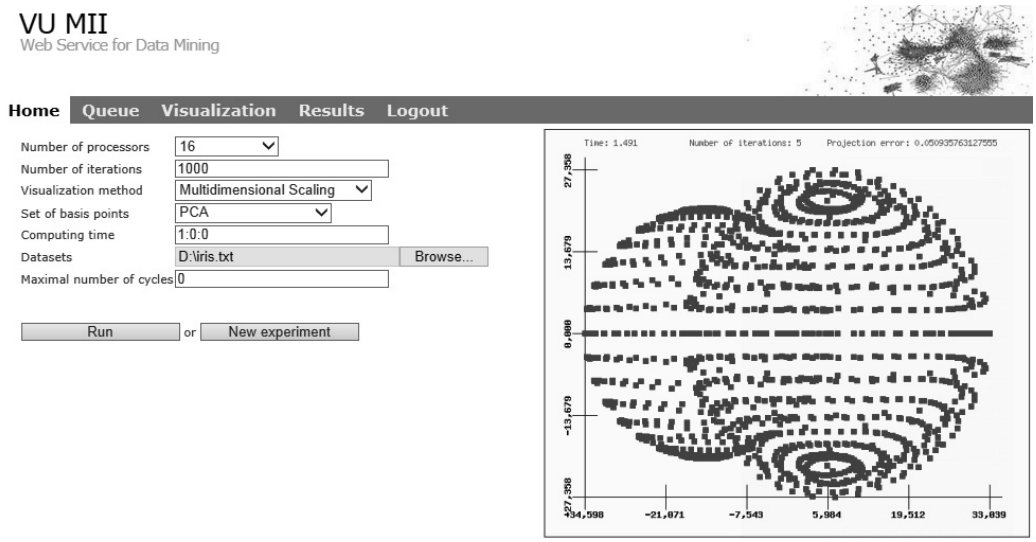
didelių tiek skaičiavimo išteklių, tiek saugojimo galimybių. Todėl pavienių kompiuterių skaičiavimo išteklių neužtenka. Didelių skaičiavimo išteklių reikalaujantiems uždaviniams spręsti naudojama vadinamoji paskirstytoji duomenų tyryba (angl. *distributed data mining*), kai duomenų tyrybos uždaviniai lygiagrečiai sprendžiami keliuose kompiuteriuose (Talia, Trunfio, 2012). Yra sukurtos kelios platformos, įgyvendinančios paskirstytąją duomenų tyrybą: *DataMiningGrid* (Stankovski et al., 2008), *Knowledge Grid* (Congiusta et al., 2008), *Discovery Net* (Čurčin et al., 2002).

Dažnai didelių skaičiavimo išteklių reikalaujantiems uždaviniams spręsti pasitelkiami gridai (angl. *Grid*). Gridas – tai laisvai prieinama, suderinta ir standartizuota infrastruktūra, kuri užtikrina lankstų, saugų, koordinuotą skaičiavimų ir informacijos saugojimo išteklių paskirstymą (Foster, Kesselman, 2004). Pagrindinis aspektas, skiriantis gridą nuo kompiuterių klasterio, yra tas, kad klasterius sudaro homogeniniai kompiuteriai, o gridą – heterogeniniai (Jadhav, 2009). Bendruoju atveju gridas – plačiai geografiškai paskirstyta infrastruktūra, kuri jungia daug skirtingo tipo išteklių. Prieigą prie šių išteklių naudotojas gali gauti iš bet kurios fizinės vietos, nepriklausomai nuo jų išdėstymo.

Gridų galimybes iliustruojančiu pavyzdžiu galima laikyti BOINC (angl. *Berkeley Open Infrastructure for Network Computing*) gridus, sukurtus naudojant programinę įrangą (BOINC, 2013). Ši programinė įranga, kurta specialiai SETI@Home (SETI@Home, 2013) projektui, kurio pagrindiniam tikslui, nežemiškos gyvybės signalų paieškai kosmoso signalų sraute, reikėjo labai didelių skaičiavimo išteklių, ilgainiui tapo atvira ir prieinama visiems norintiems kurti savo gridus. Šiuo metu BOINC programine įranga naudojasi apie 80 atvirų gridų. Pasauliniu mastu patys populiariausi – Einstein@Home (Einstein@Home, 2013) ir jau minėtas SETI@Home, gride turintys atitinkamai apie 4,2 ir 3,4 mln. kompiuterių (BOINCstats, 2013). Nors BOINC programinė įranga labai paplitusi, tačiau ji nėra vienintelė. Gridams kurti naudojamos ir kitos atvirojo kodo programos, pavyzdžiui, *Globus Toolkit* (Globus Toolkit, 2013), *GridWay* (GridWay, 2013) ir kt.

Pastaruoju metu paskirstytiems skaičiavimas vis plačiau taikoma *Hadoop* projekto programinė įranga, apimanti keletą atvirojo kodo programinės įrangos projektų, tokių kaip *MapReduce*, HDFS, *Hbase*, *Hive*, *Pig* ir kt. (Hadoop, 2012). *Hadoop* sukūrė Dougas Cuttingas, vystydamas *Apache Nutch* projektą, kuris buvo plačiai naudojamas kaip biblioteka teksto paieškai. *Apache Nutch* projektas pradėtas 2002 metais, o nuo 2006 metų tai jau *Hadoop* projektas, turintis pagrindines komponentes *MapReduce* ir HDFS. *MapReduce* – paskirstytų duomenų apdorojimo modelis ir kartu vykdymo aplinka, veikianti dideliuose kompiuterių klasteriuose (Whie, 2012). Kita svarbi komponentė – paskirstyta failų sistema, arba HDFS (angl. *Hadoop Distributed File System*). Naudojantis šia sistema, dideli failai sudalijami į 64 MB ar didesnes vienodas dalis, kurios tolygiai paskirstomos po kompiuterio klasterio kompiuterius. Palyginti su gridu, tai leidžia taupyti laiką, reikalingą duomenims persiųsti iš vieno kompiuterio į kitą. *MapReduce* – tai *Google* kompanijos inžinierių darbas, skirtas paieškos rezultatams indeksuoti, tačiau dabar ji taikoma įvairiausiose srityse: vaizdų analizei, kompiuteriniam mokymui (angl. *machine learning*). Įvairioms duomenų tyrybos sritims yra sukurta su *Hadoop* veikianti *Mahout* biblioteka (Mahout, 2011), kurioje įgyvendinti populiarūs klasifikavimo, klasterizavimo, regresijos, daugiamačių duomenų vizualizavimo (dimensijos mažinimo), evoliuciniai algoritmai. Tačiau šių algoritmų nėra daug bei dėl *MapReduce* specifikos ne visada lengvai ir efektyviai galima panaudoti esamus duomenų tyrybos algoritmus. Todėl kitame skyriuje nagrinėsime kelias populiarias duomenų tyrybos sistemas, kurios orientuojasi į saityno paslaugas. Dažnai tokios sistemos palaiko paskirstytųjų skaičiavimų galimybes.

Paskirstytosios duomenų tyrybos sistemos pradedamos kurti ir Lietuvoje. Yra sukurta interneto naršykle naudojama ir valdoma sistema, kurioje įgyvendinti daugiamačių duomenų vizualizavimo metodai (Dzemyda ir kt., 2011). Sistema leidžia vizualizuoti didelės apimties duomenų aibes (2 pav.), optimizuoti vizualizavimo algoritmų parametrus. Daugiamačių duomenų vizualizavimo metodai reikalauja daug skaičiuojamųjų sąnaudų, o rezultatus įprastai siekiama gauti per trumpą laiką, todėl visi skaičiavimai yra atliekami naudojant kompiuterių klasterį.



2 p a v. ***Daugiamačių duomenų vizualizavimo sistema***

Duomenų tyrybos sistemos

Šiame skyriuje apžvelgiamos sistemos, pagrįstos saityno paslaugomis (*Weka4WS*, *Orange4WS*, *Knime*, *Taverna*, *ClowdFlows*, DAME), kuriose įgyvendinti įvairūs duomenų tyrybos metodai.

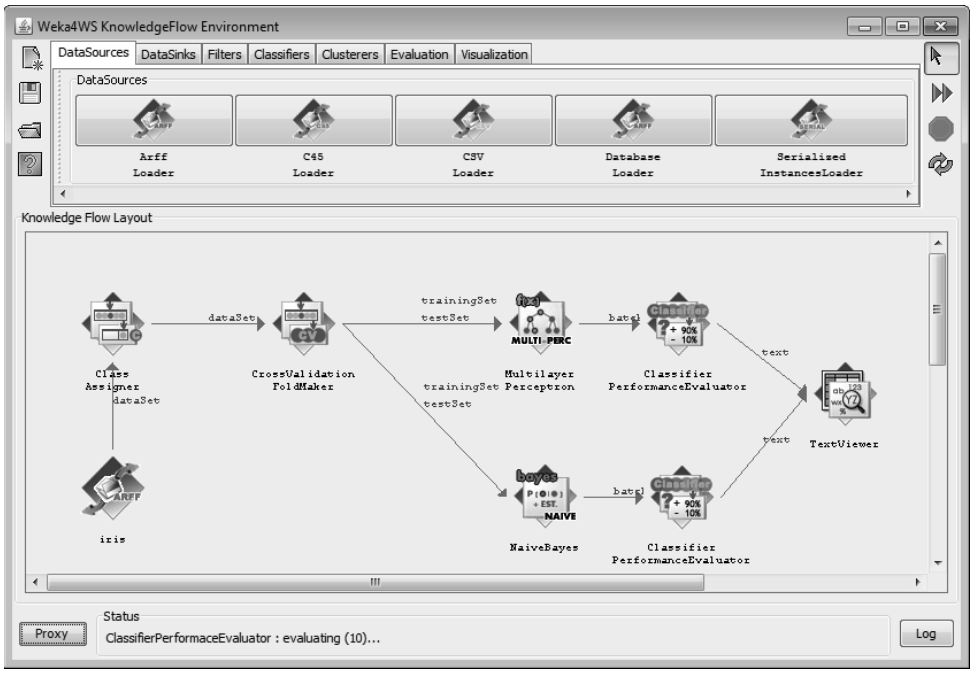
Weka4WS yra gerai žinomos populiarios *Weka* sistemos plėtinys, leidžiantis vykdyti paskirstytąją duomenų tyrybą griduose (Talia et al., 2008). Sistema turi būti įdiegta į naudotojo kompiuterį, tačiau joje įgyvendinta galimybė pasirinkti skaičiavimo išteklius ir skaičiavimus vykdyti lygiagrečiai. *Weka4WS* sukurta naudojant *Globus Aliance Java WSRF* (angl. *Web Service Resource Framework*) (Globus Toolkit, 2013) karkasą. Sistemoje yra įgyvendinta vadinamųjų darbų sekų (angl. *workflows*) sudarymo galimybė. Jos pavyzdys pateikiamas 3 paveiksle. Čia duomenys klasifikuojami dviem klasifikavimo algoritmais (*NaiveBayes* ir *Multilayer Perceptron*), kurie gali būti vykdomi skirtinguose grido mazguose.

Orange4WS yra kitos populiarios duomenų tyrybos sistemos *Orange* plėtinys (Podpecan, 2012). Lyginant su *Orange* sistema, joje įgyvendintos kelios naujos savybės: įrankiai, palengvinantys saityno paslaugų realizaciją naudojant *Python* programavimo kalbą; sistemoje įgyvendinta galimybė naudoti saityno paslaugas kaip darbų sekos komponentes.

Knime sistemoje taip pat galima naudoti saityno paslaugas (Berthold et al., 2007). Joje įgyvendintas saityno paslaugų klientas, leidžiantis prijungti jau sukurtas (išorines) saityno paslaugas prie *Knime* darbų sekų. Tam reikia pasirinkti grafinę komponentę *Generic Web Service Client*, įkelti ją į darbo langą bei nurodyti saityno paslaugos WSDL failo internetinį adresą, prijungti prie komponentės įvesties ir išvesties ar pasirinkti kitas reikalingas komponentes. Tokiu būdu bus sudaryta darbų seka, kuri esant būtinybei gali būti papildyta kitomis komponentėmis.

Programavimo ir skaičiavimo aplinkoje *Matlab* yra sukurtos funkcijos (*callSoapService*, *createClassFromWsdl*, *createSoapMessage*, *parseSoapResponse*), leidžiančios panaudoti saityno paslaugas.

Pastaruoju metu kuriamos duomenų tyrybos sistemos saityno paslaugų pagrindu. Tokie įrankiai įgyvendinti *MyGrid* projekte (myGrid, 2008), jungiančiame *myExperiment* (De Roure et al., 2008), *BioCatalogue* (Bhagat et al., 2010), *Taverna* (Missier et al., 2010) ir kitus įrankius bei aplinkas, skirtas tokioms mokslo kryptims kaip biologija, socialiniai mokslai,



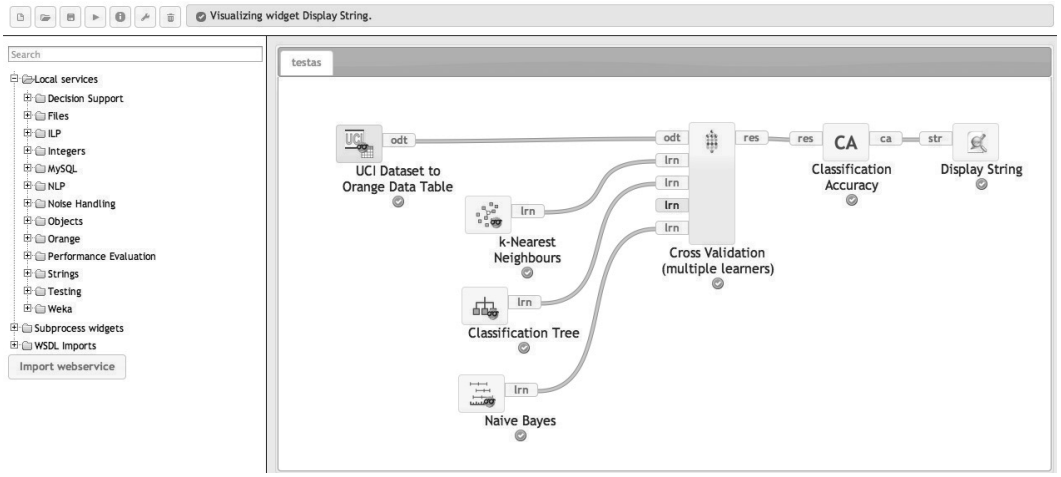
3 p a v. ***Darbų sekos pavyzdys „Weka4WS“ sistemoje***

muzika, astronomija, daugialypė terpė ir chemija. *BioCatalogu*e yra saityno paslaugų gyvybės mokslams registras. *MyExperiment* bendradarbiavimo aplinkoje mokslininkai gali saugoti savo sukurtas darbų sekas, dalytis jomis su kitais, rasti norimų.

Taverna yra atvirojo kodo ir nuo srities nepriklausanti darbų sekų valdymo sistema (Missier et al., 2010). Sistemoje nėra saityno paslaugų, skirtų duomenų tyrybai, tačiau yra galimybė išorines saityno paslaugas lengvai prijungti ir iš jų sudaryti darbų sekas. Į sistemą integruotos *myExperiment* aplinkos paslaugos bei paslaugos iš *BioCatalogue*.

Visos iki šiol apžvelgtos sistemos turi būti įdiegtos į asmeninį kompiuterį. Nauja kryptis yra duomenų tyrybos saityno programos (angl. *web applications*), kurios naudojamos ir valdomos interneto naršykle. Dažniausia programų naudotojai turi galimybę susikurti savo aplinką tyrimams ir saugoti joje gautus rezultatus. Todėl šių programų pranašumas prieš sistemas, kurias būtina įdiegti į kompiuterį, yra tas, kad naudotojas savo aplinką gali pasiekti iš bet kur interneto ryšį turinčio kompiuterio be jokio papildomo diegimo. Viena tokių programų yra *ClowdFlows* sistema (Kranjc et al., 2012). Naudotojas turi užsiregistruoti sistemoje, sukurti norimas darbų sekas ir jas išsaugoti. Sistemos vaizdas pateikiamas 4 paveiksle. Čia sudaryta tokia darbų seka: paimami duomenys, esantys UCI duomenų saugykloje (Bache, Lichman, 2013), jie klasifikuojami trimis klasifikatoriais, įgyvendintais *Orange* sistemoje, rezultatai testuojami kryžmine patikra, gauti klasifikavimo tikslumai parodomi ekrane. Sistemoje yra įgyvendintos ir *Orange*, ir *Weka* sistemose esančios saityno paslaugos. Be to, yra galimybė prijungti ir kitas saityno paslaugas, nurodant WSDL failo internetinį adresą.

DAME (angl. *DAta Mining & Exploration*) yra paskirstytosios duomenų tyrybos infrastruktūra, skirta didelės apimties duomenims analizuoti, naudojant duomenų tyrybos metodus (DAME project, 2010; Brescia et al., 2012). Ši sistema taip pat naudojama ir valdoma interneto naršykle. Skaičiavimai atliekami ne vartotojo kompiuteryje, todėl didesnių apimčių duomenų analizė gali būti atliekama daug greičiau. Į DAME sistemą



4 p a v. ***Darbų seka, sudaryta sistemoje „ClowdFlows“***

yra integruojama *Knime* sistemos funkcijos ir duomenų tyrybos metodai. DAME sistemoje neįgyvendintas darbų sekų sukūrimo mechanizmas, kuris yra daugumoje anksčiau minėtų sistemų.

Duomenų tyrybos sistemų lyginamoji analizė

Duomenų tyrybos sistemos yra lyginamos pagal įvairius kriterijus (Stankovski et al., 2008; Hmida, Slimani, 2010). Dažnai didžiausias dėmesys kreipiamas į duomenų tyrybos uždavinius, todėl trūksta lyginamosios saityno paslaugų analizės. Šiame straipsnyje atliekamos lyginamosios analizės tikslas – palyginti naujas ir vis dar tobulinamas duomenų tyrybos sistemas, pagrįstas saityno paslaugomis, bei išryškinti tų sistemų pranašumus ir trūkumus. Visos vertintos sistemos yra atvirojo kodo, laisvai prieinamos internete. Kriterijai, pagal kuriuos toliau bus lyginamos duomenų tyrybos sistemos, galimos reikšmės bei kriterijų svarbos pagrindimas, pateikiami 1 lentelėje.

Duomenų tyrybos sistemų, pagrįstų saityno paslaugomis, vertinimo rezultatai pateikti 2 lentelėje. Jei sistema atitinka nurodytą

1 l e n t e l ė. ***Sistemų lyginimo kriterijai***

Kriterijus	Galimos reikšmės	Pagrindimas
Informacijos perdavimo tarp saityno paslaugų būdas	SOAP, RESTful	Šiuo metu populiariausi būdai bendrauti su saityno paslaugomis, todėl svarbu, kad sistemose būtų įdiegiami abu būdai
Operacinės sistemos	MS Windows, Linux, Mac OS X	Svarbu, kad sistemos veiktų įvairiose operacinėse sistemose
Praplečiamumas	taip, ne	Svarbu, kad būtų įmanoma prijungti išorines saityno paslaugas be papildomo programavimo
Darbų sekos	taip, ne	Darbų sekos leidžia sukurti norimą aplinką eksperimentams, ją išsaugoti kitiems tyrimams
Saityno programa	taip, ne	Saityno programų pranašumas yra tas, kad jos naudojamos ir valdomos interneto naršykle be papildomo įdiegimo.
Duomenų tyrybos metodai	klasifikavimo, grupavimo, asociatyvių taisyklių, teksto tyryba	Sistemų universalumas yra labai svarbus duomenų tyrybos uždaviniams spręsti, nes labai dažnai tiems patiems duomenims būtina taikyti kelis skirtingus duomenų tyrybos metodus.

2 l e n t e l ė. ***Duomenų tyrybos sistemų palyginimas***

Kriterijai / Sistemos	Informacijos perdavimas		Operacinės sistemos			Praplečiamumas	Darbų sekos	Saityno programa	Duomenų tyryba				Iš viso
	SOAP	RESTful	MS Win	Linux	Mac OS X				klasifikavimo	grupavimo	asociat. taisyklių	teksto tyryba	
Weka4WS	+	–	+	+	–	–	+	–	+	+	+	–	7
Orange4WS	+	–	+	+	+	+	+	–	+	+	+	–	9
Knime	+	–	+	+	+	+	+	–	+	+	+	+	10
ClowdFlows	+	–	+	+	+	+	+	+	+	+	–	+	10
Taverna	+	+	+	+	+	+	+	–	–*	–*	–*	–*	7
DAME	–	+	+	+	+	–	–	+	+	+	–	–	7
Iš viso	5	2	6	6	5	4	5	2	5	5	3	2	

* sistemoje galima tik prijungti kitur sukurtas išorines duomenų tyrybos saityno paslaugas

kriterijų, žymima „+“, priešingu atveju – „–“. Paskutiniame stulpelyje „Iš viso“ nurodytas kiekvienos sistemos pliusų skaičius, o paskutinėje eilutėje „Iš viso“ nurodytas sistemų, atitinkančių nurodytą kriterijų, kiekis.

Weka4WS, *Orange4WS*, *Knime*, *ClowdFlows* sistemose saityno paslaugos sukurtos naudojant tik SOAP informacijos perdavimo būdą. DAME sistemoje naudojamas *RESTful*, o *Taverna* sistemoje galima įkelti ir SOAP, ir *RESTful* saityno paslaugas. Visos sistemos veikia *MS Windows* ir *Linux* operacinėse sistemose, o *Mac OS X* sistemoje neveikia tik *Weka4WS*. *Orange4WS, Knime*, *ClowdFlows* ir *Taverna* sistemose yra įgyvendinta galimybė įkelti kitur sukurtas (išorines) saityno paslaugas be papildomo programavimo. Darbų sekos įgyvendintos visose tirtose sistemose, išskyrus DAME. *ClowdFlows* ir DAME yra saityno programos. Visos keturios duomenų tyrybos metodų grupės (klasifikavimas, grupavimas, asociatyvios taisyklės bei teksto tyryba) yra įgyvendintos tik *Knime* sistemose. *ClowdFlows* sistemoje kol kas nėra įgyvendinti asociatyvių taisyklių sudarymo algoritmai. *Taverna* sistemoje nėra nė vieno duomenų tyrybos metodo, tačiau yra galimybė prijungti kitur sukurtas duomenų tyrybos paslaugas. DAME yra įgyvendinti keli klasifikavimo ir grupavimo metodai.

Pagal atliktus vertinimus didžiausiais balais įvertintos *Knime* ir *ClowdFlows* sistemos (10 iš 12 galimų). *ClowdFlows* sistemos pranašumas yra tas, kad ji yra naudojama ir valdoma interneto naršyklė. Be to ji intensyviai tebevystoma, todėl ateityje tikėtinas dar didesnis funkcionalumas.

Išvados

Pastaruoju metu nemaža dalis programinės įrangos kuriama į paslaugas orientuotos architektūros pagrindu. Ne išimtis ir duomenų tyrybos sistemos. Kelių populiarių duomenų tyrybos sistemų – *Weka* ir *Orange* pagrindu yra sukurtos saityno paslaugomis pagrįstos sistemos – *Weka4WS* ir *Orange4WS*. Dar viena populiari sistema *Knime* yra papildyta galimybe prijungti kitų kūrėjų sukurtas (išorines) saityno paslaugas. *ClowdFlows* ir DAME sistemos iš kitų išsiskiria tuo, kad sistemų nereikia įdiegti į naudotojo kompiuterį, joms naudoti ir valdyti pakanka interneto naršyklės. *Taverna* sistemoje nėra įgyvendintų duomenų tyrybos metodų, tačiau yra galimybė nesudėtingai prijungti išorines tiek SOAP, tiek *RESTful* saityno paslaugas. Visos tirtos sistemos yra atvirojo kodo, visose, išskyrus DAME sistemą, yra įgyvendinta darbų sekų konstravimo galimybė.

Darbe parinkti kriterijai, pagal kuriuos palygintos kelios saityno paslaugomis pagrįstos sistemos. Lyginamoji analizė parodė, kad pagal vertinamus kriterijus geriausiai įvertintos *Knime* ir *ClowdFlows* sistemos. Atliktos lyginamosios analizės rezultatai bus panaudoti kuriant naują duomenų tyrybos sistemą, pagrįstą saityno paslaugomis.

Padėka. Šis tyrimas atliktas Europos socialinio fondo finansuojamo projekto „Paslaugų interneto technologijų kūrimo ir panaudojimo našių skaičiavimų platformose teoriniai ir inžineriniai aspektai“ (Nr. VP1-3.1-ŠMM-08-K-01-010) lėšomis.

LITERATŪRA

BACHE, K.; LICHMAN, M. (2013). *UCI Machine Learning Repository.* University of California, Irvine, School of Information and Computer Sciences.

BERTHOLD, M.; CEBRON, N.; DILL, F.; KOTTER, T.; MEINL, T. (2007). *KNIME: The Konstanz Information Miner.* Studies in Classification, Data Analysis, and Knowledge Organization (GfKL 2007). Springer.

BHAGAT, J.; TANOH, F.; NZUOBONTANE, E.; LAURENT, T.; ORLOWSKI, J.; ROOS, M. (2010). BioCatalogue: a universal catalogue of web services for the life sciences. *Nucleic Acids Research*, W689–W694.

BOINC (2013) [interaktyvus] [žiūrėta 2013 m. liepos 8 d.]. Prieiga per internetą: <http://boinc.berkeley.edu>.

BOINCstats (2013) [interaktyvus] [žiūrėta 2013 m. liepos 8 d.]. Prieiga per internetą: <http://boincstats.com/en/stats/projectStatsInfo>.

BRESCIA, M.; LONGO, G.; CASTELLANI, M.; CAVUOTI, S.; D'ABRUSCO, R.; LAURINO, O. (2012). DAME: A Distributed Web Based Framework for Knowledge Discovery in Databases. *Memorie della Societa Astronomica Italiana Supplement,* vol. 19, p. 324–329.

CONGIUSTA, A.; TALIA, D.; TRUNFIO, P. (2008). Service-oriented middleware for distributed data mining on the grid. *Journal of Parallel and Distributed Computing*, vol. 68(1), p. 3–15.

ČURČIN, V.; GHANEM, M.; GUO, Y.; KÖHLER, M.; ROWE, A.; SYED, J. (2002). Discovery Net: towards a grid of knowledge discovery. In: *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM Press, p. 658–663.

CURK, T.; DEMŠAR, J.; XU, Q.; LEBAN, G.; PETROVIČ, U.; BRATKO, I. (2005). Microarray data mining with visual programming. *Bioinformatics*, vol. 21(3), p. 396–398.

DAME project (2010) [interaktyvus] [žiūrėta 2013 m. liepos 8 d.]. Prieiga per internetą: <http://dame.dsf.unina.it>.

DE ROURE, D.; GOBLE, C.; STEVENS, R. (2008). The Design and Realisation of the myExperiment Virtual Research Environment for Social Sharing of Workflows. *Future Generation Computer Systems*, vol. 25, p. 561–567.

DZEMYDA, G.; MARCINKEVIČIUS, V.; MEDVEDEV, V. (2011). Large-Scale Multidimensional Data Visualization: A Web Service for Data Mining. ServiceWave 2011. *Lecture Notes in Computer Science*. Springer, p. 14–25.

Einstein@Home (2013). [interaktyvus] [žiūrėta 2013 m. liepos 8 d.]. Prieiga per internetą: <http://einstein.phys.uwm.edu/>.

FAEHIM (2005). *Federated Analysis Environment for Heterogeneous Intelligent Mining* [interaktyvus] [žiūrėta 2013 m. liepos 8 d.]. Prieiga per internetą: <http://users.cs.cf.ac.uk/Ali.Shaikhali/faehim/index.htm>.

FOSTER, I.; KESSELMAN, C. (2004). *The Grid 2*. University of Southern California.

Globus Toolkit (2013) [interaktyvus] [žiūrėta 2013 m. liepos 8 d.]. Prieiga per internetą: <http://www.globus.org/toolkit>.

GridWay (2013) [interaktyvus] [žiūrėta 2013 m. liepos 8 d.]. Prieiga per internetą: <http://www.gridway.org/doku.php>.

Hadoop (2012). Apache Hadoop [interaktyvus] [žiūrėta 2013 m. liepos 8 d.]. Prieiga per internetą: <http://hadoop.apache.org/>.

HALL, M.; FRANK, E.; HOLMES, G.; PFAHRINGER, B.; REUTEMANN, P.; WITTEN, I. H. (2009). The WEKA Data Mining Software: An Update. *SIGKDD Explorations,* vol. 11(1).

HMIDA, M.; SLIMANI, Y. (2010). Meta-learning in grid-based data mining systems. *International Journal of Communication Networks and Distributed Systems*, vol. 5(3), p. 214–228.

IBM Developers Works (2013) [interaktyvus] [žiūrėta 2013 m. liepos 8 d.]. Prieiga per internetą: <http://www.ibm.com/developerworks/webservices/newto/>.

JADHAV, S. (2009). *Advanced Computer Architecture and Computing*. Technical Publications Pune.

KNELL, R. J. (2013). *Introductory R: A Beginner's Guide to Data Visualisation and Analysis using R.*

KRANJC, K.; PODPECAN, V.; LAVRAC, N. (2012). ClowdFlows: A Cloud Based Scientific Workflow Platform. *Proceedings of European Conference on Machine Learning and Knowledge Discovery in Databases*, ECML PKDD 2012. Springer, p. 816–819.

KREGER, H. (2001). *Web Services Conceptual Architecture (WSCA 1.0)*. IBM.

Mahout (2011). Apache Mahout [interaktyvus] [žiūrėta 2013 m. liepos 8 d.]. Prieiga per internetą: <http://mahout.apache.org/>.

MIERSWA, I.; WURST, M.; KLINKENBERG, R.; SCHOLZ, M.; EULER, T. (2006). YALE: Rapid Prototyping for Complex Data Mining Tasks. In: *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, p. 935–940.

MISSIER, P.; SOILAND-REYES, S.; OWEN, S.; TAN, W.; NENADIC, A.; DUNLOP, I. (2010). Taverna, reloaded. In *Proceedings of 22nd International Conference on Scientific and Statistical Database Management*, SSDBM 2010. Heidelberg, Germany.

myGrid (2008) [interaktyvus] [žiūrėta 2013 m. liepos 8 d.]. Prieiga per internetą: <http://www.mygrid.org.uk/>.

PAPAZOGLOU, M. (2003). Service-oriented computing: concepts, characteristics and directions. Web Information Systems Engineering. WISE 2003. In: *Proceedings of the Fourth International Conference on*, p. 3–12.

PODPECAN, V.; ZEMENOVA, M.; LAVRAC, N. (2012). Orange4WS Environment for Service-Oriented Data Mining. *The Computer Journal*, vol. 55(1), p. 82–98.

SETI@Home. (2013) [interaktyvus] [žiūrėta 2013 m. liepos 8 d.]. Prieiga per internetą: <http://setiathome.berkeley.edu/>.

STANKOVSKI, V.; SWAIN, M.; KRAVTSOV, V.; NIESSEN, T.; WEGENER, D., KINDERMANN, J. (2008). Grid-enabling data mining applications with DataMiningGrid: An architectural perspective. *Elsevier, Future Generation Computer Systems*, vol. 4(4), p. 259–279.

TALIA, D.; TRUNFIO, P. (2012). *Service-oriented distributed knowledge discovery*. Chapman and Hall/CRC.

TALIA, D.; TRUNFIO, P.; VERTA, O. (2008). The Weka4WS framework for distributed data mining in service-oriented Grids. *Concurrency and Computation: Practice and Experience*, vol. 20(16), p. 1933–1951.

WHIE, T. (2012). *Hadoop: The Definitive Guide.* Third Edition. Sebastopol: O'Reilly Media, Inc.

DATA MINING SYSTEMS, BASED ON WEB SERVICES

Olga Kurasova, Virginijus Marcinkevičius, Viktor Medvedev, Aurimas Rapečka

S u m m a r y

In the paper, data mining systems, based on web services, are analysed. The main notation related with web services is described. The possibilities of distributed data mining and their implementation tools – Grid, Hadoop are introduced. An analytical review of the data mining systems, based on web services, is provided. Some comparison criteria are selected. According to the criteria, a comparative analysis of the popular data mining systems, based on web services, is made. The paper illustrates, which systems are best for evaluating and which do not satisfy most of the criteria.