VILNIAUS GEDIMINO TECHNIKOS UNIVERSITETAS
MATEMATIKOS IR INFORMATIKOS INSTITUTAS

Židrina PABARŠKAITĖ

ŽINIATINKLIO ĮRAŠŲ GAVYBOS PARUOŠIMO, ANALIZĖS IR REZULTATŲ PATEIKIMO NAUDOTOJUI TOBULINIMAS

Daktaro disertacijos santrauka
Technologijos mokslai, informatikos inžinerija (07T)

Vilnius LEIDYKLA TECHNIKA 2009

Disertacija rengta 2003–2009 metais Matematikos ir informatikos institute.

Mokslinis vadovas

prof. habil. dr. Šarūnas RAUDYS (Matematikos ir informatikos institutas, technologijos mokslai, informatikos inžinerija – 07T).

Disertacija ginama Vilniaus Gedimino technikos universiteto Informatikos inžinerijos mokslo krypties taryboje:

Pirmininkas

prof. habil. dr. Gintautas DZEMYDA (Matematikos ir informatikos institutas, technologijos mokslai, informatikos inžinerija – 07T).

Nariai:

prof. dr. Romas BARONAS (Vilniaus universitetas, fiziniai mokslai, informatika – 09P),

prof. habil. dr. Jonas MOCKUS (Matematikos ir informatikos institutas, technologijos mokslai, informatikos inžinerija – 07T),

prof. dr. Dalius NAVAKAUSKAS (Vilniaus Gedimino technikos universitetas, technologijos mokslai, informatikos inžinerija – 07T),

prof. habil. dr. Rimvydas SIMUTIS (Kauno technologijos universitetas, technologijos mokslai, informatikos inžinerija – 07T).

Oponentai:

doc. dr. Regina KULVIETIENĖ (Vilniaus Gedimino technikos universitetas, technologijos mokslai, informatikos inžinerija – 07T),

prof. habil. dr. Henrikas PRANEVIČIUS (Kauno technologijos universitetas, fiziniai mokslai, informatika – 09P).

Disertacija bus ginama viešame Informatikos inžinerijos mokslo krypties tarybos posėdyje 2009 m. birželio mėn. 4 d. 15 val. Matematikos ir informatikos instituto 203 auditorijoje.

Adresas: Akademijos g. 4, LT-08663 Vilnius, Lietuva.

Tel.: (8 5) 274 4952, (8 5) 274 4956; faksas (8 5) 270 0112;

el. paštas doktor@adm.vgtu.lt.

Disertacijos santrauka išsiuntinėta 2009 m. balandžio 30 d.

Disertaciją galima peržiūrėti Vilniaus Gedimino technikos universiteto (Saulėtekio al. 14, LT-10223 Vilnius, Lietuva) ir Matematikos ir informatikos instituto (Akademijos g. 4, LT-08663 Vilnius, Lietuva) bibliotekose.

VGTU leidyklos „Technika“ 1621-M mokslo literatūros knyga.

VILNIUS GEDIMINAS TECHNICAL UNIVERSITY
INSTITUTE OF MATHEMATICS AND INFORMATICS

Židrina PABARŠKAITĖ

ENHANCEMENTS OF PRE-PROCESSING, ANALYSIS AND PRESENTATION TECHNIQUES IN WEB LOG MINING

Summary of Doctoral Dissertation
Technological Sciences, Informatics Engineering (07T)

Vilnius LEIDYKLA TECHNIKA 2009

Doctoral dissertation was prepared at the Institute of Mathematics and Informatics in 2003–2009.

Scientific Supervisor

Prof Dr Habil Šarūnas RAUDYS (Institute of Mathematics and Informatics, Technological Sciences, Informatics Engineering – 07T).

The dissertation is being defended at the Council of Scientific Field of Informatics Engineering at Vilnius Gediminas Technical University:

Chairman

Prof Dr Habil Gintautas DZEMYDA (Institute of Mathematics and Informatics, Technological Sciences, Informatics Engineering – 07T).

Members:

Prof Dr Romas BARONAS (Vilnius University, Physical Sciences, Informatics – 09P),

Prof Dr Habil Jonas MOCKUS (Institute of Mathematics and Informatics, Technological Sciences, Informatics Engineering – 07T),

Prof Dr Dalius NAVAKAUSKAS (Vilnius Gediminas Technical University, Technological Sciences, Informatics Engineering – 07T),

Prof Dr Habil Rimvydas SIMUTIS (Kaunas University of Technology, Technological Sciences, Informatics Engineering – 07T).

Opponents:

Assoc Prof Dr Regina KULVIETIENĖ (Vilnius Gediminas Technical University, Technological Sciences, Informatics Engineering – 07T),

Prof Dr Habil Henrikas PRANEVIČIUS (Kaunas University of Technology, Physical Sciences, Informatics – 09P).

The dissertation will be defended at the public meeting of the Council of Scientific Field of Informatics Engineering in the 203 auditorium of the Institute of Mathematics and Informatics at 3 p. m. on 4 June 2009.

Address: Akademijos g. 4, LT-08663 Vilnius, Lithuania.
Tel.: (8 5) 274 4952, 274 4956; fax (8 5) 270 0112;
e-mail: doktor@adm.vgtu.lt.

The summary of the doctoral dissertation was distributed on 30 April 2009.

A copy of the doctoral dissertation is available for review at the Libraries of Vilnius Gediminas Technical University (Saulėtekio al. 14, LT-10223 Vilnius, Lithuania) and Institute of Mathematics and Informatics (Akademijos g. 4, LT-08663 Vilnius, Lithuania).

Įvadas

Mokslo problemos aktualumas – dėl didėjančios konkurencijos rinkoje ieškoma naujų darbo formų, todėl didžioji dalis verslo ir ne pelno siekiančių struktūrų perkeliamos į internetinę erdvę. Tai apima įvairių tipų – įmonės-kliento, įmonės-įmonės (skirtingų verslo subjektų) bei kitokius santykius. Be to, per paskutinį dešimtmetį išaugo valstybinių institucijų, bibliotekų, asmeninių svetainių skaičius. Siūlyti prekes, teikti verslo paslaugas ar skelbti aktualią informaciją internete yra labai patogu, nes tai nepriklauso nuo geografinių ir laiko juostų skirtumų. Naudotojas, esantis kitur, nei verslo ar informacijos teikėjas, gali naršyti įmonės internetinę svetainę ir priimti sprendimą, susijusį su minėta verslo struktūra. Šis virtualus ryšys tarp tinklapių ir jų lankytojų palieka pėdsakus – įrašus arba dar kitaip vadinamus įrašus žiniatinklio žurnale, kurie kaupiasi tinklapį aptarnaujančioje tarnybinėje stotyje. Dėl tobulėjančių technologijų atsirado galimybė kaupti ir analizuoti didelių apimčių duomenis, todėl daugiau nei prieš dešimtmetį atsirado nauja tyrimų sritis – žiniatinklio įrašų gavyba. Šio žinių gavybos procesas yra panašus į kitokių duomenų (pvz. finansinių, medicininių), tačiau tam tikri šio proceso etapai yra skirtingi bei unikalūs.

Praktinė nauda, kuri gali būti gaunama analizuojant naudotojų naršymo maršrutus tinklapyje – ištirti ryšius tarp susijusių puslapių, atrasti dažniausiai pasirenkamų puslapių sekas bei tokias puslapių sekas, kurios naršomos tam tikru eiliškumu. Turint tokias žinias, galima tobulinti internetinių puslapių išdėstymo struktūrą, keisti puslapiuose esamą informaciją aktualesne, jeigu atrandama, kad tam tikra puslapių naršymo kombinacija lemia naudotojų atitinkamus veiksmus, paruošti labiausiai tikėtinus puslapius, kad, naudotojui pateikus puslapio užklausą, sutrumpėtų perduodamo į naršyklę duomenų laikas.

Darbe atliktų tyrimų metu išsiaiškinta, kad žiniatinklio įrašų filtravimui buvo skirtas nepakankamas dėmesys, nes, pašalinus nereikšmingus įrašus, duomenų analizės procesas tampa žymiai efektyvesnis. Todėl buvo sukurtas naujas duomenų filtravimo metodas, kad išgautų žinių pateikimas atitiktų tikruosius vartotojų maršrutus. Buvo nustatyta, kad paruošiant duomenis atitinkamu būdu ir suformavus fiksuoto ilgio vektorius, galima taikyti iki šiol mažai praktikoje taikytus sprendimų medžių algoritmus žiniatinklio žurnalo įrašų analizėje. O prie naudotojų žiūrėtų puslapių pridėjus ir tekstinę informaciją, esančią ant internetinių nuorodų, galima tikslinti naudotojo elgesį prognozuojančius rezultatus. Taip pat pasiūlytas rezultatų pavaizdavimo etapo tobulinimas, kuomet panaudojus tekstą, esantį ant internetinių nuorodų, rezultatai pateikiami labiau suprantama forma. Darbe atliktų tyrimų rezultatai atskleidė naujas internetinių duomenų analizės galimybes.

Tyrimų objektas. Darbo tyrimų objektas yra žiniatinklio įrašų gavyba ir su šiuo objektu susiję dalykai: žiniatinklio duomenų paruošimo etapų tobulinimas, žiniatinklio tekstų analizė, duomenų analizės algoritmai prognozavimo ir klasifikavimo uždaviniams spręsti.

Darbo tikslas. Pagrindinis mokslinio darbo tikslas – tobulinti žiniatinklio įrašų gavybos metodologiją, kuri leistų padidinti šių duomenų analizės efektyvumą.

Darbo uždaviniai. Tam, kad būtų pasiektas šis tikslas, darbe reikėjo išspręsti sekančius uždavinius: 1. Atlikti žiniatinklio duomenų literatūros apžvalgą, pateikti šios mokslo srities taksonomiją, paaiškinti duomenų šaltinius bei tipus. 2. Išanalizuoti žiniatinklio įrašų duomenų ruošimo etapus, susisteminti egzistuojančius metodus. 3. Sukurti duomenų filtravimo metodą, kuris sumažintų nereikalingų analizei įrašų kiekį. 4. Pateikti duomenų paruošimo būdą, leidžiantį žinių gavybai patogiai ir efektyviai panaudoti sprendimų medžius. 5. Panaudoti tekstą, esantį ant internetinių nuorodų ir ištirti, kaip jis įtakoja žiniatinklio duomenų prognozavimo uždavinių tikslumą. 6. Ištirti, kaip žiniatinklio analizės duomenis tyrėjui pateikti labiau suprantama semantine forma.

Tyrimų metodika. Atlikti analitiniai tyrimai nagrinėjant internetinių duomenų gavybos literatūrą. Apžvelgti žinių gavybos algoritmai, taikomi žiniatinklio įrašų analizėje. Tyrimo metu buvo remtasi informacija apie tinklapių kūrimo metodus, kad atkreipti dėmesį, kaip svetainių struktūra įtakoja žiniatinklio įrašų kiekį. Analizuojant žiniatinklio naudotojų elgesio šablonus, buvo naudojami sprendimų medžių, dirbtinių neuroninių tinklų bei artimiausio kaimyno metodai. Gautų išvadų patikimumui įvertinti naudoti standartiniai matematikos statistikos metodai. Eksperimentams buvo naudojami egzistuojančių internetinių svetainių duomenys. Pasiūlyti metodai buvo realizuoti naudojant Borland C++ Builder 6, Microsoft Visual C#. Microsoft Access ir SQL Server buvo naudoti talpinti duomenis, atlikti užklausas ir kurti ataskaitas.

Mokslinis naujumas

1. Išryškintas poreikis kurti žiniatinklio žurnalo įrašus filtruojantį universalų metodą.
2. Pasiūlyti būdai, kaip paruošti duomenis, kad būtų galima taikyti sprendimų medžius klasifikavimo uždaviniams.

3. Pasiūlyta ir ištirta strategija, kaip pagerinti lankytojų elgesį prognozuojančius rezultatus naudojant ne tik istorinę informaciją, bet ir tekstą.

4. Pasiūlytas būdas kaip informacija žiniatinklio įrašų tyrėjui būtų pateikta suprantamesne forma.

Praktinė vertė. Disertacijoje atskleisti duomenų paruošimo būdai leidžia atlikti žiniatinklio įrašų tyrimus panaudojus tobulesnius analizės būdus. Detaliai išnagrinėtas žiniatinklio įrašų paruošimo etapas, išryškintos silpnosios vietos bei pasiūlyti nauji metodai atveria pažangesnes technologines galimybes šios mokslo krypties specialistams. Pasiūlytų duomenų filtravimo ir rezultatų pateikimo metodų efektyvumas atsiskleistų jas diegiant programinėje įrangoje analizuojančioje lankytojų naršymo ypatumus. O darbe pasiūlytais technologiniais patobulinimais bei rezultatais, gautais naudojant sprendimų medžius bei kitais klasifikatoriais, galėtų pasinaudoti svetainių projektuotojai, tinklalapių kūrėjai bei žinių gavybos srities ekspertai.

Ginamieji teiginiai

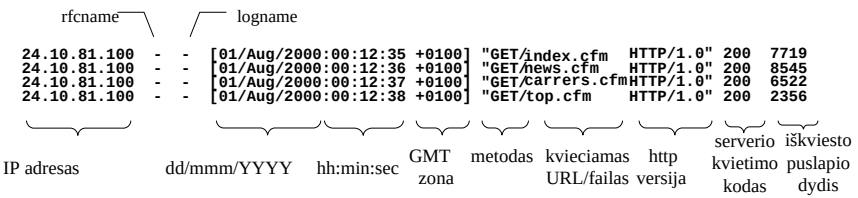
1. Naujas žiniatinklio įrašų filtravimo metodas, paliekantis įrašus, atitinkančius tikruosius naudotojų maršrutus, kelis kartus (priklausomai nuo svetainės dizaino) sumažina nereikšmingų duomenų kiekį, todėl tiksliau paaiškina sprendimo algoritmų šablonus.
2. Duomenų paruošimas, kurio metu suformuojami fiksuoto ilgio vektoriai, leidžia panaudoti sprendimo medžius prognozavimo uždaviniams spręsti, kas padeda efektyviai atpažinti ir paprastai paaiškinti tinklalapio naudotojų elgseną.
3. Klasifikavimo uždavinių tikslumą padidina požymių vektoriuje kartu su naudotojų naršytų puslapiais panaudojus juos žyminčią semantinę informaciją.
4. Pasiūlytas nuorodų teksto panaudojimas rezultatų pavaizdavimo etape leidžia tyrėjui matyti rezultatus aiškesne forma.

Darbo apimtis. Disertacija parašyta anglų kalba, darbo apimtis – 136 puslapių teksto su priedais, 32 paveikslai, 15 lentelių, panaudotas 184 literatūros šaltinis.

1. Žiniatinklio duomenų gavyba

Šiame skyriuje analizuojamas žiniatinklio žurnalo įrašų žinių gavybos etapai: duomenys ir jų ruošimas, žinių gavyba ir rezultatų pateikimas.

Žurnalo įrašai būna įvairių formatų, tačiau, pats populiariausias yra taip vadinamas „įprastas žurnalo formatas“ (1 pav., 1 lentelė).

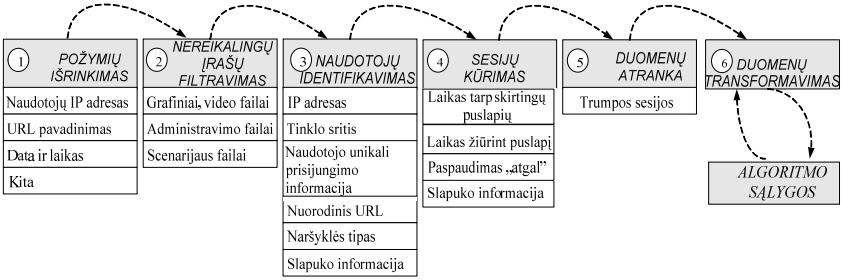


1 pav. Įprastinis žurnalo įrašų žiniatinklio serveryje pavyzdys

1 lentelė. Žurnalo įrašų laukų paaiškinimai

Laukas	Paaiškinimas
Skaitinis kompiuterio adresas arba IP adresas	Kompiuterio, prijungto prie interneto adresas, sudarytas iš skaičių, vienareikšmiškai identifikuojantis kompiuterį internete
rfcname	Naudotojo identifikacija
logname	Naudotojo slaptažodis
dd/mmm/YYYY	dd – diena, mmm – mėnuo, YYYY – metai
hh:min:sec	hh nurodo valandą, min – minutes, sec – sekundes
GMT zona	Laiko zona pagal Grinvičą
Metodas	Naudotojo kvietimo tipas
Kviečiamas URL	Naudotojo kviečiamas tinklalapis
HTTP versija	Protokolo versija, kuriuo parsiunčiama informacija į naudotojo naršyklę
Serverio kvietimo kodas	Nurodo kvietimo sėkmės ar klaidos kodą
Iškviesto puslapio dydis baitais	Baitų kiekis, persiunčiamas iš serverio į naudotojo naršyklę (kompiuterį)

Prieš atliekant duomenų analizę, žiniatinklio įrašus reikia paruošti, t.y. išrinkti tik tam tikrus požymius, išfiltruoti nereikalingus įrašus, atpažinti unikalius tinklalapio naudotojus, suformuoti unikalių tinklalapio naudotojų sesijas bei paruošti duomenis pagal metodo, kuriuo bus atliekama analizė, reikalavimus. Detali žiniatinklio žurnalo įrašų paruošimo schema pavaizduota 2 paveiksle.



2 pav. Žiniatinklio įrašų paruošimo etapai

Žiniatinklio žurnalo įrašų analizėje plačiausiai naudojami klasterizavimo, asociacijų taisyklių bei sekų algoritmai.

Klasterizavimas. Šio metodo principu vyksta naudotojų apjungimas į grupes pagal panašias charakteristikas. Panašumas tarp skirtingų naudotojų skaičiuojamas naudojant atstumo funkciją. Jis dažniausiai būna *Euklido* ir skaičiuojamas pagal formulę (1):

$$d_E(i,i+1) = \left(\sum_{i=1}^{p} (x_k(i) - x_k(i+1))^2\right)^{\frac{1}{2}}, \quad (1)$$

čia n – naudotojų skaičius, p – matavimų skaičius, aprašantis naudotojus, i ir $i+1$ yra naudotojai duomenyse, tai i-tojo taško vektorius apibrėžiamas formule (2):

$$x(i) = (x_1(i), x_2(i), ..., x_k(i), ..., x_p(i)),\ 1 \le i \le n,\ 1 \le k \le p. \quad (2)$$

Klasterizavimas plačiai taikoma atvejais, kai istorinė informacija apie lankytojų elgesį panaudojama tam, kad pritaikyti svetainę tam tikrų grupių naudotojų poreikiams. Todėl tokiai analizei dažnai panaudojama ir kita informacija, pvz. žiniatinklio struktūra, naudotojų asmeniniai duomenys.

Asociacijų taisyklės. Kadangi asociacijų taisyklės atpažįsta susijusius elementus, objektus ar veiksmus, kurie įvyksta vieno įvykio metu, jos plačiai naudojamos puslapių analizėje. Asociacijų taisyklių algoritmas pateikia rezultatus tokia forma: *A=> B(S, c)*, kai *A* ir *B* yra lankytojo naršytų puslapių vienoje sesijoje rinkinys. Kadangi rinkinių variantų kiekis gali būti labai didelis, taisyklių kokybei kontroliuoti įvedami tam tikri dydžiai – kriterijai. Taikant šiuos kriterijus – dažnumą (*S*) ir patikimumą (*c*), sumažinamas taisyklių skaičius, kurių dauguma būna nereikšmingos ir labai retai pasitaikančios.

Dažnumas skaičiuojamas pagal formulę (3):

$$S = P(A \cup B) = \frac{\text{sesijos turinčios A ir B}}{\text{visos sesijos}}. \quad (3)$$

Patikimumas skaičiuojamas pagal formulę (4):

$$c = \frac{P(A \cup B)}{P(A)} = \frac{\text{sesijos turinčios A ir B}}{\text{sesijos turinčios A}}. \quad (4)$$

Sekų taisyklės. Panašiai, kaip ir asociacijų, sekų taisyklės pateikia panašius rezultatus, bet jie yra paremti eiliškumu. Taikant sekų taisykles, svarbiausias faktorius yra nustatyti maksimalų taisyklių kiekį, kurių sekos yra ilgiausios pagal užsiduotą dažnumą.

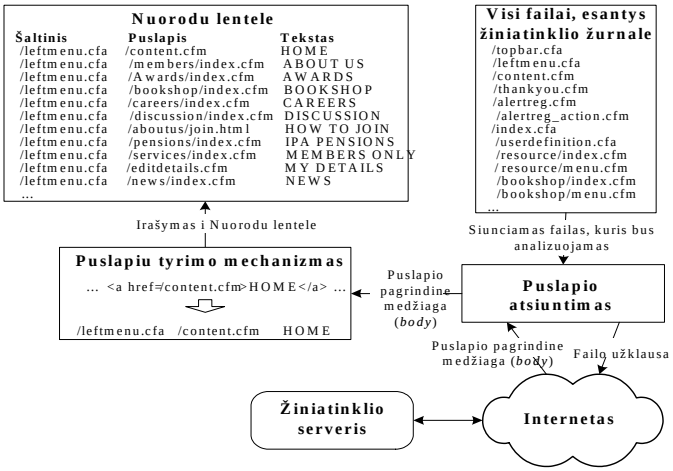
Žiniatinklio žurnalo įrašų žinių pateikimui naudojamos dvimatės diagramos ir grafikai, grafų teorijos principu paremtos vizualizacijos priemonės.

2. Svetainių kūrimo struktūros

Šiame skyriuje atlikta HTML protokolo bei pagrindinių HTML ženklinimo kalbų, naudojamų rašyti tinklalapius, apžvalga bei lyginamosios koncepcijos. Naudojant HTTP protokolą, žiniatinklio serveriai pateikia savo išteklius (svetaines, tinklalapius) lankytojų interneto naršyklėse. Dėl žiniatinklio svetainių kūrimo būdų įvairovės, žiniatinklio serveris kaupia skirtingus įrašus. Kodėl taip yra, kaip vyksta procesas, kurio metu ištekliai yra perduodami naršyklei – tai aptariama šiame skyriuje, o detaliai nagrinėjama paskutinėse skyriaus dalyse.

3. Nuorodomis paremtas filtravimo metodas

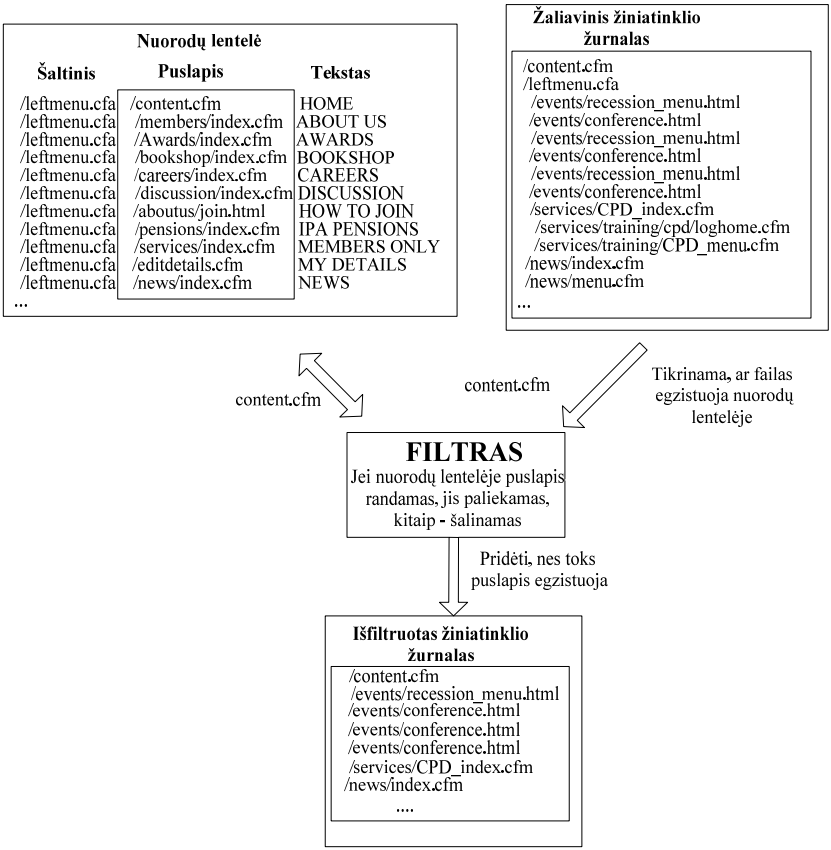
Šiame skyriuje pateikiamas autoriaus pasiūlytas naujas žiniatinklio žurnalo įrašų filtravimo metodas. Pirmas šio metodo etapas – puslapio hipertekstinės informacijos parsisiųsdinimas iš žiniatinklio žurnalo įrašų lauko, nusakančio naudotojo kviestus failus (puslapius) sudaroma lentelė.



3 pav. Puslapių tyrimo mechanizmo schema

Po to įvykdoma programa, kuri kiekvieną tos lentelės įrašą siunčia žiniatinklio serveriui ir užklausia failo hipertekstinę informaciją. Gavęs užklausą, serveris grąžina puslapio turinį puslapių tyrimo mechanizmui, šis išnagrinėja gautą hipertekstinę informaciją, išrenka nuorodas pagal specialų nuorodų atpažinimo elementą <a>… ir įrašo jas į *nuorodų lentelės* lauką *puslapis* (3 pav.). Ši procedūra pakartojama visiems žiniatinklio žurnalo failams. Žurnalo įrašai atitinka nuorodų lentelės *šaltinis* lauką, o įrašai laukelyje *puslapis* – tai nuorodos, esančios atitinkamame puslapyje. Pastarieji įrašai ir yra faktiškos tinklapyje egzistuojančios nuorodos. Pvz., iš žurnalo įrašų paimamas failas */leftmenu*, iškviečiamas jo turinys. Išrenkamos visos nuorodos iš puslapio */leftmenu*, kurios patalpinamos nuorodų lentelės laukelyje *Puslapis: /contemt.cfm, /members/index.cfm, /Awards/index.cfm ir t.t.*

Antrasis etapas – filtravimas. Šiame etape šalinami nereikšmingi įrašai. Nereikšmingas įrašas apibrėžiami failai, esantys žiniatinklio žurnale, tačiau kurių turinio negalima išgauti naršyklėje atlikus „pelės paspaudimo operaciją”. Tokie failai-įrašai yra kitų puslapių sudėtinės dalys. Todėl puslapio hipertekstinė informacija negali būti gauta ir puslapio tyrimo mechanizmas neturi ką nagrinėti. Filtravimo mechanizmas veikia sekančiai: paimamas įrašas iš žurnalo ir tikrinama, ar jis yra nuorodų lauke (lentelė *nuorodų lentelė*, laukas *puslapis*). Nerastas *nuorodų lentelėje* įrašas pašalinamas. Jei įrašas randamas, jis yra įrašomas į naują lentelę *išfiltruotas žiniatinklio žurnalas* (4 pav.).

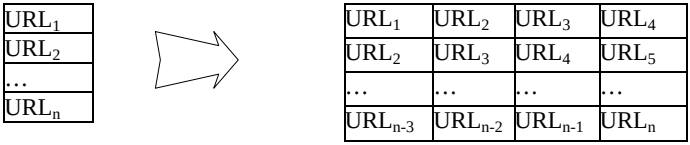


4 pav. Nuorodų išrinkimas

4. Žiniatinklio įrašų žinių gavyba naudojant sprendimų medžius

Kadangi sprendimo medžiai objektų charakteristikas pateikia taisyklių pavidalu, kurios suprantamos daugeliui vartotojų, jie dažnai naudojami klasifikavimo/prognozavimo uždaviniams spręsti. Skyriuje aprašyta duomenų paruošimo technologija, kuri reikalinga naudojant C4.5 medžio algoritmą. C4.5 buvo pasirinktas iš grupės algoritmų, nes mokymo procesas yra greitas, suprantamas rezultatų interpretavimas. Duomenys turėjo būti paruošti taip, kad atitiktų klasifikatoriaus reikalavimus. Vienas iš reikalavimų – fiksuoto ilgio duomenų vektoriai. Tačiau apsilankymų internetinėje svetainėje metu lankytojų

parsisiųsdintų puslapių skaičius yra skirtingas, t. y. jis nėra fiksuotas. Naudojant vektorius, susidedančius iš ilgų sekų, pvz., iš 6-ių puslapių, sugeneruotos taisyklės turi labai mažai atvejų (žemi dažnumo ir patikimumo rodikliai), pvz., 3 atvejai iš 1000, o naudojant trumpas sekas, pvz., iš 2-jų puslapių, taisyklės tampa per daug bendros (pvz., jei „pradinis puslapis", tai „pardavimai") ir netikslios. Todėl, kaip optimalus variantas, buvo nuspręsta analizuoti tokias taisykles, kurios susideda iš 4-ių puslapių. Tam tikslui buvo naudojamos sesijos iš 4-ių ir daugiau puslapių ir daromi „lango" tipo įrašai, t. y. kad vienas vektorius būtų sukonstruotas iš 4-ių puslapių. Taigi, buvo imami įrašai nuo 1-o iki 4-o, po to kitas to paties naudotojo vektorius sudaromas iš įrašų nuo 2-o iki 5-o, nuo 3-čio iki 6-o, nuo 4-o iki 7-o ir t. t., kol baigiasi vieno lankytojo įrašų seka vieno apsilankymo svetainėje metu (5 pav.). Kadangi sprendimų medžiai gali dirbti tik su fiksuotu skaičiumi klasių, buvo nuspręsta puslapius apjungti juos apibendrinant. Eksperimentinio tinklalapio puslapiai buvo sugrupuoti pagal tam tikras temas.



5 pav. Duomenų vektoriaus formavimas

Pvz.: *location.php?id=1, location.php?id=5, location.php?id=21* buvo priskirti vienai grupei *location*. Neatlikus šio apibendrinimo, gaunasi be galo didelis skaičius klasių ir sprendimų medžius taikyti tampa nevertinga.

Turint tokį duomenų vektorių buvo suformuluoti tokie praktiniai uždaviniai:

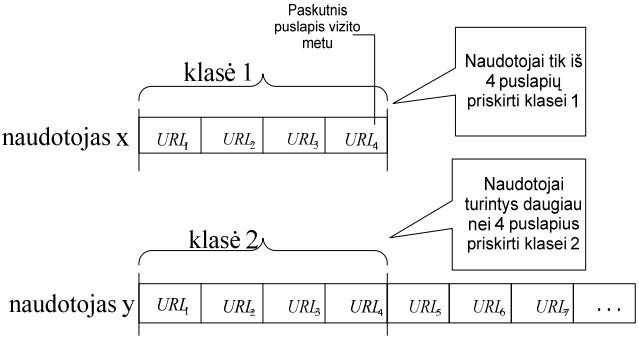
1. žinant 3-is puslapius, prognozuoti 4-ą puslapį,
2. surasti 4 pradinius puslapius, po kurių naudotojas baigia naršyti,
3. surasti 4 pabaigos puslapius, po kurių naudotojas baigia naršyti.

1 Eksperimentas:

Prognozuojamas 4-as puslapis žinant 3 prieš tai einančius puslapius.
Jei S(i)=URL1 **ir** S(i+1)=URL2 **ir** S(i+2)=URL3 **tai** S(i+3)=URL4 **pabaiga,**
Kur i yra puslapis sesijos metu S. Taisyklių pavyzdžiai:
95% apsilankiusių /dininganddancing.cfm, /nightlife.cfm ir /location.cfm taip pat domėjosi /onelocation.cfm. Tai sudarė 0,05% visų vizitų (42 iš 42750).
87% apsilankiusių /events.cfm, /jobmarket.cfm, /classifieds.cfm taip pat domėjosi /oneclass.cfm. Tai sudarė 0,15% nuo visų vizitų (128 iš 42750).

2 Eksperimentas:

Klasifikuojami puslapiai, po kurių naršymas baigiamas, o po kurių tęsiamas, imant 4 pradinius puslapius (6 pav.).

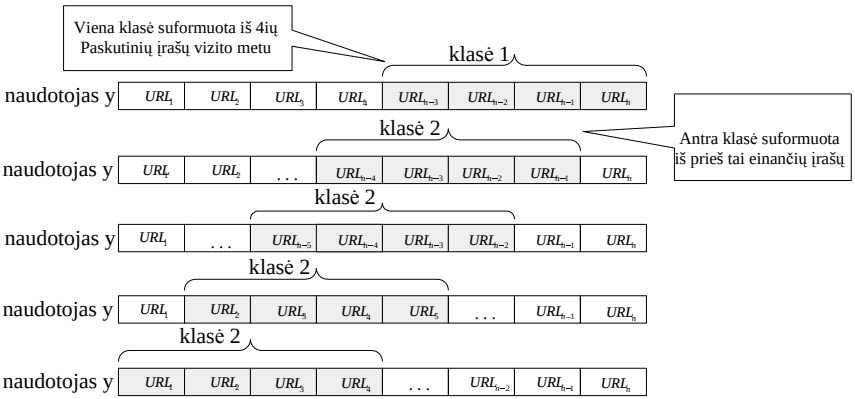


6 pav. Duomenų ruošimas 2-am eksperimentui

Taisyklės pavyzdys: **Jei** S1 = /classifieds.cfm **ir** S2 = /oneclass.cfm **ir** S3 =/events.cfm **ir** S4=/todayall.cfm **tai baigia naršyti.** Tai sudarė 2,27% visų sesijų (907 iš 39991).

3 Eksperimentas:

Klasifikuojami puslapiai, po kurių baigiamas naršymas, o po kurių tęsiamas, imant 4 pabaigos puslapius (7 pav.).



7 pav. Klasių ruošimas 3-iajam eksperimentui

Taisyklės pavyzdys: **Jei** S(i)=/events.cfm **ir** S(i+1)=/onelocation.cfm **ir** S(i+2)=/leisure.cfm **ir** S(i+3)=/kids.cfm **tai** pabaiga**.** Tai sudarė 0,26% nuo visų tos klasės vizitų (19 iš 7258).

5. Tekstas žiniatinklio žurnalo įrašų analizėje

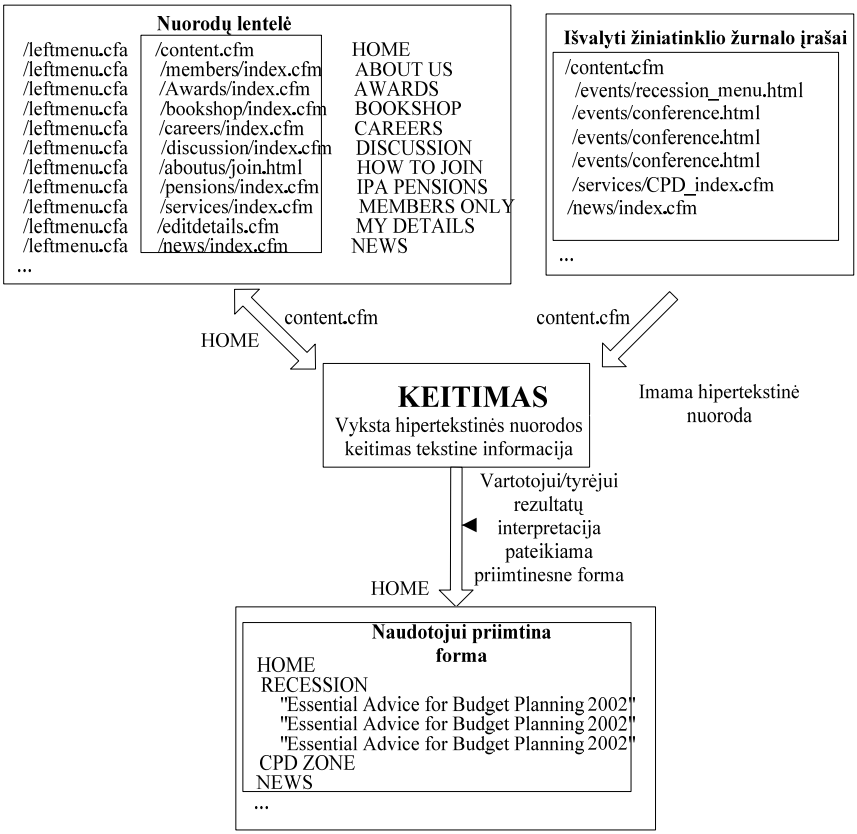
Šiame skyriuje pasiūlytas metodas kartu su žiniatinklio žurnalo įrašų duomenimis naudoti ir tekstą. Pasiūlyto metodo naujovė – naudojamas tekstas, esantis ant internetinių nuorodų. Todėl, ruošiant duomenis, kiekvieną žiniatinklio įrašą reikia apjungti su tekstu. Žemiau pateikti etapai, kaip tekstas yra ruošiamas:

- parsisiųsdinamas kiekvieno puslapio hipertekstinės informacijos turinys,
- turinys analizuojamas, išrenkama tekstinė informacija, kuri apibrėžiama specialiais simboliais <a>…,
- tekstinė informacija išnagrinėjama lingvistiškai, iš žodžių sudaromas sąrašas, skaičiuojama, kiek kartų kiekvienas žodis pasikartojo sąraše,
- galutiniame etape buvo sudarytas fiksuoto ilgio pusiau dvejetainis vektorius, t. y. jeigu puslapis ar tekstas neegzistavo sesijos metu, įrašomas nulis, kitaip prie esamo dydžio pridedamas vienetas.

Pasiruošimo eksperimentams santrauka:

1. naudotas fiksuoto ilgio vektorius. Požymiai: žiūrėtų puslapių skaičius, sesijos trukmė, 119 puslapiai, 126 žodžiai (viso – 247 požymiai),
2. duomenų imtys: vienasluoksniui ir daugiasluoksniui perceptronui atvejams – mokymo (33%), tikrinimo (33%), testinė (33%), artimiausio kaimyno atveju – mokymo (66%) ir testinė (33%),
3. normalizavimas – vienasluoksniui ir daugiasluoksniui perceptronui, artimiausio kaimyno,
4. daugiasluoksnis perceptronas – trijų sluoksnių, 3 paslėpti neuronai vidiniame sluoksnyje. Įėjimų skaičius buvo lygus požymių skaičiui, t. y. 2+119+126, o išėjimų – klasių skaičiui, t. y. 2.
5. eksperimentas 1: klasifikuoti lankytojus į grupes (registruotus ir svečius, imant tik abejoms grupėms pasiekiamus tinklapius). Klasės: 8160 – svečiai, 8408 – registruoti naudotojai.
6. eksperimentas 2: prognozuoti naudotojų elgesį. Ar lankytojas sugrįš į svetainę? Klasės: 12312 – sesijų, kurias suformuoja naudotojai apsilankę svetainėje daugiau nei 1-ą kartą, 4255 – apsilankę tik vieną kartą.
7. eksperimentai kartoti 5 kartus kiekvienu algoritmu, galutinė klaida – visų eksperimentų klaidų vidurkis.

Kita sritis, kur buvo panaudotas nuorodų tekstas – tai rezultatų pateikimo etape. Metodo esmė paremta tuo, kad hipertekstinės nuorodos keičiamos tekstu, kurį paprastai lankytojai mato ant tinklalapio hipertekstinių nuorodų (8 pav.). Pasiūlius atlikti šį pakeitimą, patobulinamas rezultatų pateikimas duomenų tyrėjui.



8 pav. Procesas, kaip hipertekstinės nuorodos keičiamos tekstu, esančiu ant nuorodų

Bendrosios išvados

Atlikus mokslinės literatūros analizę, pasiūlius ir išnagrinėjus naujas duomenų paruošimo bei rezultatų pateikimo išraiškas, išryškinus naujas žiniatinklio įrašų analizės galimybes ir atlikus kompiuterinius modelių skaičiavimus, suformuluotos šios mokslinės bei praktinės išvados:

1. Atlikta analitinė internetinių duomenų analizės literatūros apžvalga. Išnagrinėti ir susisteminti internetinių duomenų analizės etapai: duomenų filtravimas (nereikalingų įrašų šalinimas), duomenų paruošimas (naudotojų identifikavimas, sesijų sukūrimas), algoritmų taikymas ir rezultatų pateikimas duomenų tyrėjui.

2. Sukurtas ir patikrintas žiniatinklio įrašų filtravimo metodas. Atlikus metodo įvertinimą nustatyta, kad jis pašalina žymiai daugiau nereikalingų įrašų tipų bei „suklysta“ rečiau priskiriant reikalingus analizei įrašus netinkamų įrašų grupei.

3. Pasiūlyta taikyti tokius duomenų paruošimo būdus, kurie leistų panaudoti sprendimų medžių algoritmą C4.5 internetinių svetainių žiniatinklio įrašų analizei. Ekspertas gali lengvai ir suprantamai vertinti ryšius tarp skirtingų puslapių.

4. Pasiūlytas jungtinis metodas žiniatinklio įrašų klasifikavimo uždaviniams su vienasluoksniu ir daugiasluoksniu perceptronu bei artimiausio kaimyno algoritmais kaip požymius naudoti ne tik žiniatinklio žurnalo įrašus, bet ir internetinės svetainės nuorodų tekstą. Geriausi rezultatai buvo gauti naudojant vienasluoksnį perceptroną. Klaida sumažėjo 3,1% (nuo 20,8% iki 17,6%). Sprendžiant prognozavimo uždavinį, geriausias rezultatas pasiektas naudojant daugiasluoksnį perceptroną. Klaidos dydis sumažėjo 3,5% (nuo 23,6% iki 20,1%).

5. Sukurtas metodas leidžia analizės rezultatus pateikti naudojant realios internetinės svetainės semantinę išraišką, kuri yra labiau suprantama nei techninės internetinės nuorodos.

Autoriaus mokslinių publikacijų disertacijos tema sąrašas

Straipsniai recenzuojamuose mokslo žurnaluose

1. Pabarskaite, Z.; Raudys, A. 2007. A process of knowledge discovery from web log data: Systematization and critical review. *Journal of Intelligent Information Systems*. 28(1): 79–104. ISSN 0925-9902. (Thomson ISI Web of science)

2. Raudys, S; Pabarskaite, Z. 2004. Fixed Non-linear Combining Rules versus Adaptive Ones, *in Lecture Notes in Computer Science. Springer-Verlag*. 260–265. ISSN 0302-9743. (Thomson ISI Web of science)

3. Pabarskaite, Z. 2003. Decision trees for web log mining. *Intelligent Data Analysis*. 7(2): 141–154. ISSN 1088-467X.

Straipsniai kituose leidiniuose

4. Pabarskaite, Z.; Raudys, A. 2002. Advances in Web usage mining, in *Proceedings of the 6th World Multiconference on Systemics, Cybernetics and Informatics*. 11. 508–512. ISBN 980-07-8150-1.

5. Pabarskaite, Z. 2002. Implementing advanced cleaning and end-user interpretability technologies in Web log mining, in. *Proc. of the 24th International Conf. on Information Technology Interfaces (ITI 2002)*. 1. 109–113. ISSN 1330-1012.

Trumpos žinios apie autorių

Židrina Pabarškaitė gimė, užaugo ir mokėsi Vilniuje.

1997 m. įgijo informatikos inžinerijos bakalauro laipsnį Vilniaus Gedimino technikos universiteto Fundamentinių mokslų fakultete. 1999 m. įgijo informatikos inžinerijos magistro laipsnį Vilniaus Gedimino technikos universiteto Fundamentinių mokslų fakultete. 2003–2009 – Matematikos ir informatikos instituto doktorantė. 1997 m. stažavosi ABB Švedijoje, 1999–2003 m. dirbo duomenų analizės srityje Jungtinėje Karalystėje, Londone.

ENHANCEMENTS OF PRE-PROCESSING, ANALYSIS AND PRESENTATION TECHNIQUES IN WEB LOG MINING

Topicality of the problem – Internet is becoming an important part of our life; therefore more attention is paid to the information quality on the web and how it is displayed to the user. This knowledge can be extracted by gathering web servers' data – log files, where all users' navigational patters are recorded. The research area of this work is web log data analysis in order to enhance information presentation on the web. Web log data analysis steps are similar to other kind of data analysis (e. g. financial, medical) but some processes are different and unique.

The research objects of the dissertation are web log data cleaning methods, data mining algorithms and web text mining. The key aim of the work is to improve pattern discovery steps mining web log data in order to:

1. improve the quality of the data for researchers who analyse users behaviour,
2. improve the ways how information is presented, to speed up information display to the end user.

Reseach object. The research object of the dissertation is web log data mining process. General topics that are related with this object: web log data preparation methods, data mining algorithms for prediction and classification tasks, web text mining. The key target of the thesis is to develop methods how to improve knowledge discovery steps mining web log data that would reveal new opportunities to the data analyst.

The aim of the work– the aim of this research is to improve the effectiveness of currently available web log mining systems by proposing innovative cleaning, analysis and results presentation methods.

Tasks of the work. In order to obtain those goals, the following tasks have to be done: 1. Provide a comprehensive web log mining (and related to web design fields) literature review. Identify web design peculiarities. Systemize which ones influence types of files collected by web servers. 2. Propose efficient data cleaning method with minimum information loss required for subsequent data analysis. 3. Overview techniques for mining web log usage data. Provide a practical study of effects and limitations using decision trees for prediction tasks. 4. Develop an integrated web log and web content data mining framework to model various real world situations and provide a study which could lead to a better prediction and classification quality. 5. Investigate how results in the hypertext format could be presented to the data analyst in the semantically more understandable format.

Methodology of research. Theoretical and empirical analysis, comparison of known data mining and web mining methods was performed. Also knowledge from text retrieval area was used. Methods of web page design have been reviewed in order to understand the organisation of the files which are recorded into web server log file. Real world data was used for experimental study. C++ and C# languages where used for testing hypothesis and implementing proposed methodologies, Microsoft Access and SQL server databases were used for storing data, running queries and generating reports.

Scientific novelty

1. Performed systemized review of methods used for web log mining. Investigated existing web log data pre-processing, analysis and results presentation methods. Methods are classified, systemized and referred to a relevant web log mining analysis steps. On the basis of this theoretical investigation, it was established that data pre-processing takes a majority of

time in knowledge discovery process and influences analysis stage by reducing number of records and analysis time.

2. In the study about different design structures it was showed that the amount of data gathered by web server depends on web pages design. To remove redundant data, a new data cleaning method has been introduced. The proposed cleaning framework enables to view only actual visitors‘ clicks.

3. It is demonstrated that decision tree approach can be used with reasonable misclassification error for analysing navigational users’ patters and generated sequential pages resulting in browsing termination or continuing browsing.

4. Introduced combined approach which takes users browsing history and text appeared on the links for mining web log data. Proposed methodology increased accuracy of various prediction tasks.

5. Cognitive aspects of web designers’ and end users’ allowed proposing more understandable way for displaying web log mining analysis results.

Practical value. Theoretical material and developed models can contribute to the research community for improving web log data mining tools. Web site administrators and web developers could use enhancements proposed in these thesis, which have been revealed applying decision trees and other classification techniques.

Defended propositions

1. Proposed cleaning technique reduces number of records and makes analysis process faster. Moreover, “link based” technique imitates real clicks therefore easier to trace visitors navigational patterns in the results examination phase.

2. Performing specific data transformation and constructing fixed length vectors allows using decision trees for web site visitor behaviour analysis.

3. Experimental evaluation using not only visitors navigational, but textual information as features increase classification accuracy.

4. Perception and interpretation of the results becomes clearer and more attractive because they appear as a text, which users see while browsing the actual web site.

The scope of the scientific work. Thesis consist of introduction, 5 chapters, conclusions, references, list of author‘s publications and 3 appendixes.

The content

Chapter 1 gives an introduction to the data mining and process to extract knowledge from the data. Here background theory of web logs, web mining taxonomy and data sources available for analysis is presented. Stages of the knowledge discovery process analysed. However, due to the way web log data is collected, many unclear issues exist. For example: what records are relevant, how unique user must be defined, how users browsing sessions must be identified and etc. Thus, theoretical concepts and variety of different techniques are presented in this chapter which deals with these kinds of problems. Finally, web usage analysis and visualisation/examination techniques are presented with a brief references to the other research works. A lot of literature sources are referred throughout the chapter. Chapter 2 describes how data transfer protocol works, introduces to the concepts and processes occurring between the web and users who download web documents. Depending on the web site type, different files are recorded in web log (where users' accessed pages are recorded). All these issues discussed in this chapter have to be considered in order for reader to understand the delicacy of web data preparation process and the need for a proper cleaning mechanism. Chapter 3 presents an implementation of a new technique for the web log data cleaning. The chapter discusses about problems occurred cleaning web log data. A comparison study and examples of different cleaning techniques gives a detail view on this subject. Chapter 4 presents an approach how to organise specific data preparation in order to use decision trees in web log mining. Chapter 5 presents a new methodology for web usage mining, which combines text tags as new features. The aim of this chapter is to show effectiveness of using such methodology solving various data mining tasks. An approach is presented to perform technical software outcome to the artefact which are human attractive and well accepted by the business people.

General conclusions

1. An investigative study was undertaken to determine essential characteristics of web mining. Analysis was performed on a selection of web usage tools and methodologies proposed by research community. Examination number of web usage tools allowed compare results based on a set of standard techniques and ones proposed in these thesis.

2. Systematisation of existing web log cleaning techniques was done. Author introduced new web log data record cleaning method. The method performs "link based" cleaning. This enables web log data practitioners to use only essential information for further data analysis process.

3. Specific data preparation process was proposed in order to compose fixed length vectors. This enables executing various prediction tasks and allows understanding users' behaviour using decision tree algorithms.

4. Introduced combined approach which takes users browsing history and text appeared on the hyperlinks. Combined approach helps better understand and predict users' behaviour. In the task to classify users into groups, combined (pages and text) method increased accuracy by 3.1% (error rate from 20.8% decrease to 17.6%) . In the task "will user returns to the site", combined (pages and text) method increased accuracy by 3.5% (error rate from 23.6% decrease to 20.1%).

5. Proposed more understandable approach for displaying web log mining results. In this approach instead of hyperlink we display text associated with that hyperlink.

Short description about the author of the dissertation

Židrina Pabarškaitė was born, grew up and studied in Vilnius.

First degree in Informatics Engineering, Faculty of Fundamental Sciences, Vilnius Gediminas Technical University, 1997. Master of Science in Informatics Engineering, Faculty of Fundamental(Colin 2004) Sciences, Vilnius Gediminas Technical University, 1999. 2003–2008 PhD student at the Institute of Mathematics and Informatics. 1997 was on internship at ABB Sweden. In 1999–2003 was working in academic and commercial environment related to data mining, United Kingdom, London.

Židrina Pabarškaitė

ŽINIATINKLIO ĮRAŠŲ GAVYBOS PARUOŠIMO, ANALIZĖS IR REZULTATŲ PATEIKIMO NAUDOTOJUI TOBULINIMAS

Daktaro disertacijos santrauka
Technologijos mokslai, informatikos inžinerija (07T)

Židrina Pabarškaitė

ENHANCEMENTS OF PRE–PROCESSING, ANALYSIS AND PRESENTATION TECHNIQUES IN WEB LOG MINING

Summary of Doctoral Dissertation
Technological Sciences, Informatics Engineering (07T)

2009 04 21. 15,0 sp. l. Tiražas 100 egz.
Vilniaus Gedimino technikos universiteto
leidykla „Technika“, Saulėtekio al. 11, LT-10223 Vilnius,
http://leidykla.vgtu.lt
Spausdino UAB „Biznio mašinų kompanija“,
J. Jasinskio g. 16A LT-01112 Vilnius