

DUOMENŲ GAVYBOS TECHNOLOGIJOS NUOTOLINIŲ STUDIJŲ KURSŲ DUOMENŲ BAZĖS ANALIZEI

Donatas Tamašauskas, Tomas Bersėnas, Sigita Turskienė

Šiaulių universitetas, Informatikos katedra

Įvadas

Šiuolaikinės kompiuterinės technologijos užtikrina didelių duomenų srautų gavimą ir jų saugojimą. Tačiau vis dar lieka nemaža spraga tarp duomenų surinkimo bei saugojimo ir jų suvokimo. Šią spragą padeda užpildyti duomenų gavyba. Pasak Gartner Group, duomenų gavyba yra prasmingų šablonų, dėsningumų, modelių ir tendencijų radimo procesas dideliuose informacijos kiekiuose, naudojantis modelių atpažinimo, statistiniais bei matematiniais metodais (Bernatavičienė, 2008; Sakalauskas, 2009).

Duomenų gavybos technologijų taikymas švietime gali padidinti galimybę planuoti studijų veiklą, atskleisti proceso dalyvių elgesį ir pan. Tai naudinga ne tik nuotolinių studijų procesui efektyvinti, bet ir padėtų prognozuoti naudojamų kompiuterių serverių resursų kiekį, kuriuose būtų patalpintos aptarnaujančios sistemos (Stravinskienė, Gudas, 2010).

Duomenų gavybos technologijų taikymą švietime tyrinėjo G. Dzemyda, P. Gudynas, V. Saltenis, V. Tiesis (2001), atlikdami mokytojų kvalifikacijos, amžiaus kaitos bei jų skaičiaus dinamikos tyrimą, taikant dirbtinius neuronų tinklus. S. Turskienė, G. Kulvietis, R. Burbaitė (2003) – atlikdami Jaunųjų kompiuterininkų mokyklos vartotojų duomenų analizę, Saulius Preidys, Leonidas Sakalauskas (2011) – mokymosi stilių nustatymą ir vizualizavimą, analizuojant studentų veiklą.

Lietuvoje gana plačiai paplitusi atvirojo kodo Moodle virtuali mokymosi aplinka (VMA), leidžianti patalpinti virtualius studijų kursus internetinėje terpėje, ir dėl to bet kuriuo laiku, iš bet kurios vietos įmanoma pasiekti ne tik reikiamą informaciją ar atlikti įvairias užduotis, bet ir bendrauti su kitais sistemos vartotojais.

Tyrimo tikslas – nuotolinių studijų kursų duomenų bazių analizė duomenų gavybos technologijomis.

Uždaviniai: sukurti programinę įrangą, susisteminančią VMA Moodle sistemų duomenų bazių duomenis; programa STATISTICA 7 atlikti gautos duomenų bazės duomenų gavybą.

Tyrimo metodai. Mokslinės ir techninės literatūros analizė, duomenų analizė.

Institucijos padalinio VMA Moodle sistemų duomenų bazių analizė

Analizuojami duomenis yra vieno iš institucijos padalinio, kuris iš viso naudoja keturiolika VMA

Moodle sistemų, kurios daugiausia yra 1.9.x versijos, kelios yra 2.x versijos.

Moodle sistemų vartotojai suskirstyti į tris grupes: dėstytojai, studentai ir administratoriai. Kiekvienos grupės vartotojas turi institucijos vidinį kodą, kuris yra vienodas visose Moodle sistemose.

Šiuo metu visi Moodle sistemų vartotojai yra kuriami ar atnaujinamos jų paskyros, kai gaunami duomenys iš kito institucijos padalinio. Tačiau Moodle sistemose dar galima rasti vartotojų, kurių paskyros buvo sukurtos kitais būdais (elektroniniu paštu, be patvirtinimo).

Visiems aktyviems vartotojams priskirti vaidmenys, kurie visose Moodle sistemų duomenų bazėse yra vienodi, t. y. jų identifikavimai su vaidmens pavadinimu sutampa, tačiau gali būti, kad ateityje ir nesutaps, į tai reikės atsižvelgti kuriant Moodle sistemų duomenų bazių sisteminančią programinę įrangą. Pagal vartotojo vaidmenį bus galima filtruoti, kurie vartotojai, kuriai vartotojų grupei priklauso.

Administratorių grupės vartotojai savo paskyras naudoja visose Moodle sistemose tik administravimo tikslams. Šie vartotojai, norėdami kurti dalykus, privalo turėti dėstytojų grupės vartotojus.

VMA Moodle sistemų duomenų bazės susisteminti programinė įrangą

Remiantis institucijos padalinio VMA Moodle sistemų duomenų bazių analizės rezultatais, sukurtas programinis įrankis PHP programavimo kalba, panaudojant diagramų braižymo biblioteką *JpGraph* ir *HTML Table Filter Generator* biblioteką, skirtą HTML lentelės dinaminiam filtravimui.

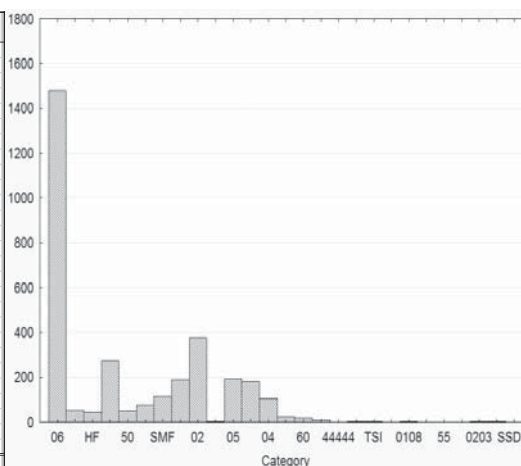
Ši programinė įrangą susistemina duomenis iš neriboto kiekio Moodle duomenų bazių, kurios gali būti 1.9.x ar 2.x versijų. Realizuoti dviejų tipų – administratoriaus teises ir peržiūros teises turintys – vartotojai. Ši programinė įrangą atvaizduoja duomenis pagal pasirinktą ataskaitą iš visų nurodytų duomenų, kuriuos galima išsaugoti CSV failo formatu bazių. Programinės įrangos ataskaitos: kiek ir kokius studijų dalykus studijuoja kiekvienas studentas, visi studijuojami dalykai, dėstytojai ir jų dėstomi dalykai, dalykų aktyvumas pagal studentus, dalykų aktyvumas pagal dėstytojus, studentai ir jų veiksmai VMA, vartotojų prisijungimo kiekiai pagal pasirinktą laiko intervalą. Kiekvieną ataskaitą pagal pasirinktus kriterijus galima analizuoti atskirai ir analizės rezultatus

galima išsaugoti CSV failo (= rinkmenos) formatu arba, jei numatyta galimybė, atvaizduoti stulpelinėje ar skritulinėje diagramoje.

Susistemintos Moodle sistemų duomenų bazės duomenų gavyba

Duomenims analizuoti buvo panaudotos VMA

Category	Count	Cumulative Count	Percent	Cumulative Percent
06	1481	1481	46,05100	46,0510
SGNSF	53	1534	1,64801	47,6990
HF	45	1579	1,39925	49,0983
01	274	1853	8,51990	57,6182
50	50	1903	1,55473	59,1729
07	76	1979	2,36318	61,5361
SMF	114	2093	3,54478	65,0808
13	190	2283	5,90796	70,9888
02	375	2658	11,66045	82,6493
EF	4	2662	0,12438	82,7736
05	193	2855	6,00124	88,7749
03	182	3037	5,65920	94,4341
04	105	3142	3,26493	97,6990
MIF	26	3168	0,80846	98,5075
60	20	3188	0,62189	99,1294
GMF	11	3199	0,34204	99,4714
44444	1	3200	0,03109	99,5025
NENURODYTA	2	3202	0,06219	99,5647
TSI	2	3204	0,06219	99,6269
1672	1	3205	0,03109	99,6580
0108	2	3207	0,06219	99,7201
MF	1	3208	0,03109	99,7512
55	1	3209	0,03109	99,7823
0182	1	3210	0,03109	99,8134
0203	2	3212	0,06219	99,8756
SLK	3	3215	0,09328	99,9689
SSDG	1	3216	0,03109	100,0000
Missing	0	3216	0,00000	100,0000



1 pav. Fakultetų dažnių metodo apdoravimo rezultatai

Nustatyta, kad fakultetas „06“, t. y. socialinių mokslų fakultetas, turi didžiausią skaičių studentų grupės vartotojų. Dažnių lentelėje pastebima, kad šio fakulteto studentų grupės vartotojai sudaro net ~46 % visų studentų grupės vartotojų (1 pav.).

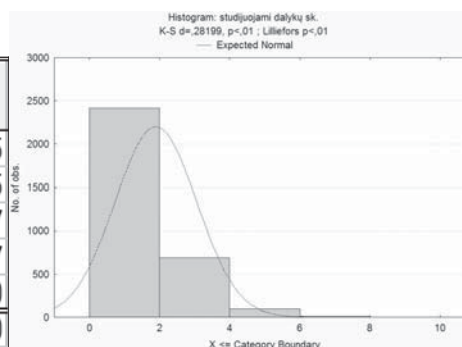
Studentų grupės vartotojų studijuojamų daly-

Moodle sistemų duomenų bazių sisteminančios programinės įrangos ataskaitos.

Duomenys apie studentų grupės vartotojus ir jų studijuojamus dalykus analizuoti taikant dažnių metodą. Atlikta studentų, studijuojančių fakultetuose, dažnių analizė (1 pav.).

kų kiekiui analizuoti pasitelktas aprašomosios statistikos metodas. Tyrimo rezultatai parodė (2 pav.), kad studentų grupės vartotojai daugiausia studijuoja po 2 dalykus, net ~75 %. Tai puikiai atsispindi studentų grupės vartotojų studijuojamų dalykų kiekio dažnių lentelės stulpelinėje diagramoje (2 pav.).

Category	Count	Cumulative Count	Percent of Valid	Cumulative % of Valid	% of all Cases	Cumulative % of All
0,000000<x<=2,000000	2417	2417	75,15547	75,1555	75,15547	75,1555
2,000000<x<=4,000000	688	3105	21,39303	96,5485	21,39303	96,5485
4,000000<x<=6,000000	97	3202	3,01617	99,5647	3,01617	99,5647
6,000000<x<=8,000000	11	3213	0,34204	99,9067	0,34204	99,9067
8,000000<x<=10,00000	3	3216	0,09328	100,0000	0,09328	100,0000
Missing	0	3216	0,00000		0,00000	100,0000



2 pav. Studentų grupės vartotojų studijuojamų dalykų kiekio dažnių metodo apdoravimo rezultatai

Aprašomosios analizės būdu gauti rezultatai (3 pav.) parodė, kad studijuojamų dalykų vidurkis yra ~1,89 ir kad pusė visų studentų grupės vartotojų studijuoja po 2 dalykus, net 1620 iš 3216 analizuoja-

mų studentų grupės vartotojų, o daugiausia studentų grupės vartotojai studijuoja po dešimt dalykų. Bendras studijuojamų dalykų skaičius yra 6092.

Variable	Valid N	Mean	Median	Mode	Frequency of Mode	Sum	Minimum	Maximum
studijuojami dalykų sk.	3216	1,894279	1,000000	1,000000	1620	6092,000	1,000000	10,00000

3 pav. Aprašomoji analizė studentų studijuojamų dalykų kiekiu

Šių metų duomenų apie studentų grupės vartotojus ir jų aktyvumą analizei taikytas klasterizavimo k-vidurkių metodas. Nagrinėjant duomenis pagal 5 klasterius, pastebėta, kad daugiausia studentų pate-

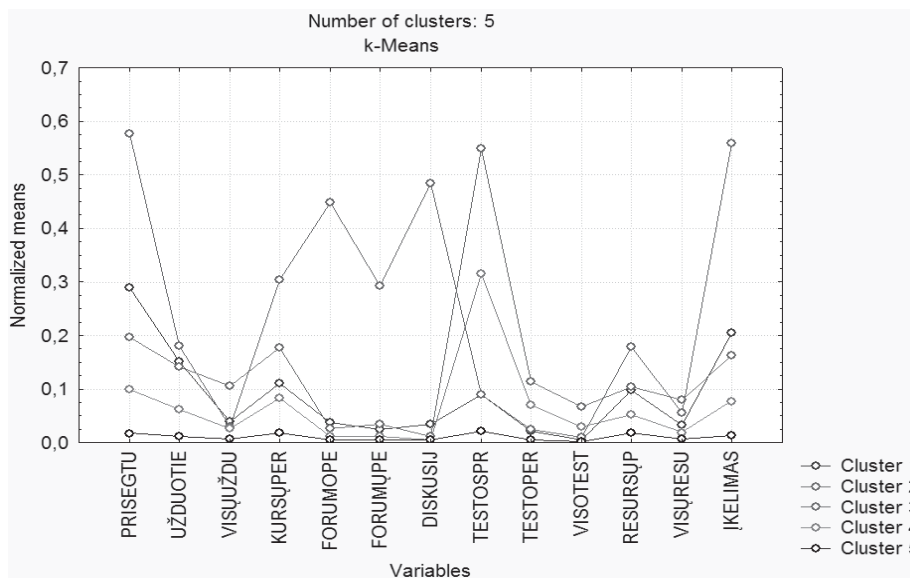
ko į penktąjį klasterį, kuris pagal pateiktos lentelės (4 pav.) aktyvumo kategorijų vidurkius, lyginant su kitų klasterių aktyvumo kategorijų vidurkiais, yra žemiausias.

Cluster	PRISEGTU	UŽDUOTIE	VISUŽDU	KURSUŠPER	FORUMOPE	FORUMUPE	DISKUSIJ	TESTOSPR	TESTOPER	VISOTEST	RESURSŪP	VISŪRESU	IKELIMAS	Number of cases	Percentage(%)
1	12,44681	83,32766	9,13191	240,0638	28,1064	3,24681	27,5277	4,39149	15,65532	1,41277	193,4383	3,910638	13,14468	235	7,35985
2	8,48583	78,40081	24,33603	384,5101	19,1093	4,51417	9,6275	26,87449	82,53846	16,08907	206,0283	9,829960	10,37247	247	7,73567
3	24,77778	99,96296	6,88889	659,7407	335,8148	38,29630	382,3333	4,37037	18,07407	2,44444	356,1111	6,740741	35,77778	27	0,84560
4	4,24870	34,26943	5,93782	179,1244	8,5959	1,58290	4,5803	15,42487	51,17358	7,04145	102,2798	2,313472	4,95855	386	12,08894
5	0,74587	6,67450	1,42037	40,7855	4,0440	0,69843	3,5601	1,02785	3,97041	0,63577	35,7711	0,869887	0,84987	2298	71,96993

4 pav. Klasterių lentelė

Todėl galima teigti, kad studentų grupės vartotojai yra pasyvūs, nes jie sudaro ~72% visų studijuo-

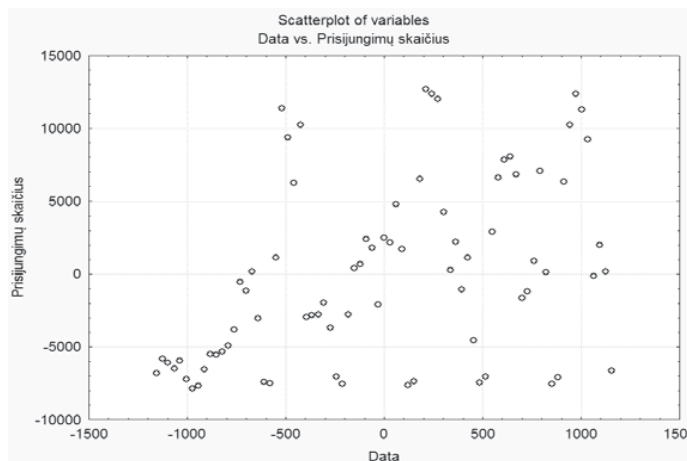
jančių studentų (5 pav.).



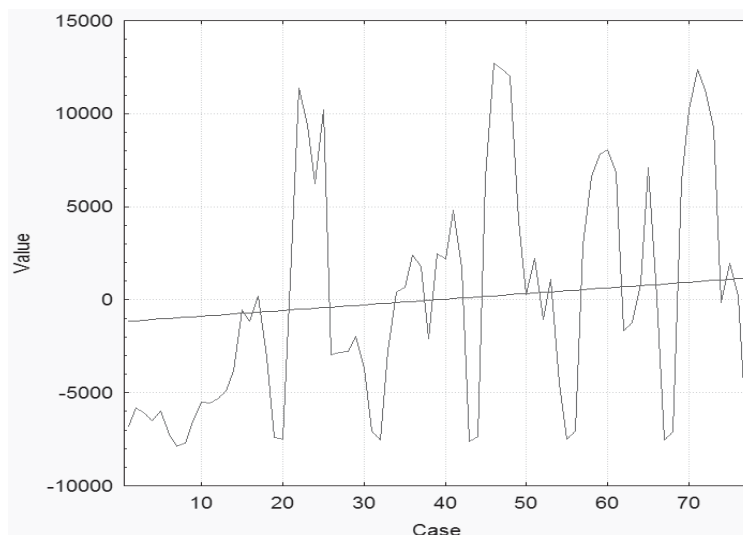
5 pav. Klasterių pasiskirstymo grafikas

Analizuojant nuo 2006 metų iki 2012 metų gegužės mėnesio visų Moodle sistemų vartotojų prisijungimo kiekius, kiekvieną mėnesį buvo taikytas nepriklausomų komponentų analizės metodas (angl. *independent component analysis*). Studentų grupės

vartotojų kiekio pagal laiko intervalą taškų sklaida pavaizduota 6 pav. Remiantis studentų grupės vartotojų prisijungimo grafiku (7 pav.), nustatyta, kad studentų prisijungimų kiekis nuo 2006 metų nuolat augo.



6 pav. Studentų grupės vartotojų kiekio pagal laiko intervalą taškų sklaida



7 pav. Studentų grupės vartotojų prisijungimų grafikas

Susistemintos duomenų bazės pildymo problemos ir jų sprendimo būdai

1. Kai kurie studentų grupės vartotojai turi dėstytojų grupės teises, ši problema buvo išspręsta papildžius studentų grupės vartotojų filtravimą pagal tai, kad studento grupės vartotojų prisijungimo varde būtų du taškai, nes studentų grupės vartotojų prisijungimo vardų struktūra yra „metai.pavardė.vardas“.
2. Pastebėta, kad visose Moodle duomenų bazėse dalykų pirmas įrašas yra Moodle sistemos domeno vardas. Todėl, rašant užklausa, buvo atsižvelgta į tai, kad įtraukiant visus studijuojamus dalykus iš visų duomenų bazių, nebūtų nepaimti pirmi įrašai.
3. Atliekant analizę pagal studijų dalyko kodą, išaiškinta problema – dėstytojų grupės vartotojai neteisingai pildo informaciją apie savo studijų dalyką. Pavyzdžiui, ne visi dėstytojų grupės vartotojai vienodai įrašo į dalyko informacijos kortelę dalyko kodą ir numerį, todėl dalyko analizė pagal dalyko kodą tapo negalima.
4. Analizuojant duomenis, nustatyta, kad dalies studentų vartotojų vienodos grupės ar fakultetai priskirti mažosiomis raidėmis, kitų – didžiosiomis raidėmis. Problema išspręsta – prieš analizuojant duomenis, jų įrašuose mažąsias raides pakeisti didžiosiomis.
5. Analizuojant studentų grupes, pastebėta, kad vieniems vartotojams prie grupės parašyta „nenurodyta“, kitiems – tuščias laukelis. Ši problema buvo išspręsta pakoreguojant užklausas, kad atkreiptų dėmesį į šią problemą ir priskirtų grupes nenurodytai kategorijai.
6. Analizuojant dėstytojų grupės vartotojus, paaiškėjo – kai kurie iš jų naudojami paskyromis, kuriuose nenurodyti vidiniai kodai, todėl negalima

atlikti jų rūšiavimo. Tai padaryta pagal prisijungimo vardą, tačiau susidurta su trimis problemomis: 1) pastebėta, kad vieno vartotojo pavardė vienoje duomenų bazėse be lietuviškų raidžių, kitose diferencijuota, pvz., e ar é; 2) pastebėta, kad vienas iš vartotojų vienoje Moodle sistemoje kuria dalykus, naudodamasis dėstytojo grupės vartotojo paskyra, o kai kuriose kitose sistemoje – administratoriaus grupės vartotojo paskyra, kurių prisijungimo vardai skiriasi; 3) filtruojant pagal prisijungimo vardą, dėstytojų grupės vartotojų analizė tampa netiksli pagal prisijungimo būdą. Atsižvelgiant į iškilusias problemas, nuspręsta nedaryti išsamesnės dėstytojų grupės vartotojų analizės.

7. Studentų grupės vartotojų nurodytos grupės kartais nurodomos kartu su pogrupiu. Nurodymai būna įvairūs: per brūkšnį, pasviręs brūkšnys į kairę ar į dešinę ir pan. Dėl šios priežasties teko atsisakyti išsamesnės analizės pagal studentų grupės vartotojų grupes.

Susistemintos duomenų bazės pildymo rekomendacijos

1. Nustatyti dėstytojų grupės vartotojų taisykles, kaip teisingai pildyti studijų dalykų korteles.
2. Sutvarkyti skriptus, kuriais duomenys perduodami į Moodle duomenų bazę, kad nebeliktų vienuodų grupių ar fakultetų, tačiau su skirtingų dydžių raidėmis. Šio dalyko sprendimas taip pat padėtų išspręsti lietuviškų raidžių vartojimo problemą.
3. Išspręsti vartotojų duomenų atnaujinimą, kad informacija atsinaujintų ne tuomet, kai vartotojai prisijungia, o tada, kai gaunama informacija apie vartotojus iš kito institucijos padalinio.
4. Dėstytojų grupės vartotojams drausti naudotis paskyromis, kurioms nepriskirti vidiniai kodai.

5. Sugriežtinti tvarką dėl administratoriaus grupės vartotojų paskyrų naudojimo ne *Moodle* sistemų administravimo darbams.
6. Išspęsti problemą dėl pogrupių nurodymo grupės laukelyje, vienas iš problemos sprendimo būdų – suderintas prefiksas, pagal kurį būtų nurodomas pogrupis.

Išvados

1. Dėl iškilusių problemų, kuriant programinę įrangą susisteminančią VMA *Moodle* sistemų duomenų bazių duomenis, nebuvo galima numatyti visų institucijos *Moodle* sistemų duomenų bazių išsamios analizės.
2. Atliekant duomenų analizę, paaiškėjo, kad daugiausia VMA *Moodle* naudojami socialinių mokslų fakulteto studentai. Pastebėta, kad institucijoje nemažai studentų studijuoja po 2 studijų dalykus,

be to, studentų prisijungimo skaičius kiekvieną mėnesį nuo 2006 metų nuolat augo.

Literatūra

1. Bernatavičienė J., 2008, *Vizualios žinių gavybos metodologija ir jos tyrimas*. Vilnius.
2. Dzemyda G., Gudynas P., Saltenis V., Tiesis V., 2001, *Lietuvos pedagogai ir moksleiviai: analizė ir prognozė*. Vilnius.
3. Preidys S., Sakalauskas L., 2011, *Nuotolinio mokymosi stilių personalizavimas. Informacijos mokslai*. Vilnius.
4. Sakalauskas L., 2009, *Duomenų gavyba*. Vilnius.
5. Stravinskienė A., Gudas S., 2010, *Duomenų gavybą paremta veiklos modelių*. Kaunas.
6. Turskienė S., Kulvietis G., Burbaitė R., 2003, Duomenų gavybos technologijų panaudojimas jaunųjų kompiuterininkų mokyklos veiklos analizei. *Lietuvos matematikos rinkinys*. T. 43. Spec. nr. P. 297–303.

DATA MINING TECHNOLOGIES FOR ANALYSIS OF DATABASES FOR DISTANCE COURSES

Donatas Tamašauskas, Tomas Bersėnas, Sigita Turskienė

Summary

Nowadays, distance studies are becoming one of the most popular learning methods in Lithuania and the whole world. Therefore it is important to analyze the activities performed during the process of distance studies. Data mining technologies help to analyze the process of studies, the activities of its participants, the behaviour of the learners, etc.

After completing the analysis of the Moodle database of Šiauliai University the data processing software for the University's virtual learning environment Moodle was created based on its results. This software is used to receive various data from Moodle database, which are analyzed in different aspects by using data mining technologies.

Keywords: data mining, data analysis, databases.

DUOMENŲ GAVIMO TECHNOLOGIJOS NUOTOLINIŲ STUDIJŲ KURSŲ DUOMENŲ BAZĖS ANALIZEI

Donatas Tamašauskas, Tomas Bersėnas, doc. Sigita Turskienė

Santrauka

Šiuo metu nuotolinės studijos tampa vienu populiariausių studijų būdų Lietuvoje ir pasaulyje, todėl svarbu atlikti nuotolinių studijų proceso veiklų analizę. Duomenų gavybos technologijos padeda analizuoti studijų procesą, jo dalyvių veiklas, besimokančiųjų elgesį ir pan.

Atlikta Šiaulių universiteto *Moodle* duomenų bazės analizė, kurios rezultatais remiantis sukurta Universiteto virtualios mokymosi aplinkos *Moodle* duomenų apdorojimo programinė įranga. Šia programine įranga gaunami įvairūs *Moodle* duomenų bazių duomenys, kurie analizuojami duomenų gavybos technologijomis keliais pjūviais.

Prasminiai žodžiai: duomenų gavimas, duomenų analizė, duomenų bazės.

Įteikta 2012-05-17