KAUNO TECHNOLOGIJOS UNIVERSITETAS
INFORMATIKOS FAKULTETAS
INFORMACIJOS SISTEMŲ KATEDRA

Martynas Ruzgys

IT žinių portalo statistikos modulis pagrįstas grupavimu

Magistro darbas

Darbo vadovas
prof. dr. R. Butleris

Kaunas, 2007

KAUNO TECHNOLOGIJOS UNIVERSITETAS
INFORMATIKOS FAKULTETAS
INFORMACIJOS SISTEMŲ KATEDRA

Martynas Ruzgys

IT žinių portalo statistikos modulis pagrįstas grupavimu

Magistro darbas

Recenzentas
doc. dr. V. Pilkauskas
2007-05-28

Vadovas
prof. dr. R. Butleris
2007-05-28

Atliko
IFM-1/2 gr. stud.
Martynas Ruzgys
2007-05-28

Kaunas, 2007

TURINYS

1 ĮVADAS 3
2 STATISTIKOS MODELIŲ IR METODŲ ANALIZĖ 5
2.1 Duomenų gavyba ir grupavimas 5
2.2 Euklido kvadratinės paklaidos matas 7
2.3 K-vidurkių grupavimo metodas 7
2.4 Duomenų grupavimo taikymas statistikos sistemoms 10
2.4.1 Mineset (SGI) – duomenų gavybos ir analizės sistema 10
2.4.2 STATISTICA Data Miner (StatSoft) – duomenų gavybos, analizės, prognozavimo sistema 14
2.4.3 StarProbe Data Miner (Rosella) – duomenų gavybos, analizės, prognozavimo sistema 15
2.4.4 Internetinių portalų sprendimai 17
2.5 Siūloma portalo duomenų analizės sistema 19
3 STATISTIKOS MODULIO KONCEPTUALIOJI SPECIFIKACIJA 20
3.1 Funkcionalumas 20
3.1.1 Veiklos kontekstas 20
3.1.2 Sistemos panaudojimo atvejai 21
3.1.3 Funkcionalumo pjūviai 22
3.2 Duomenų vaizdas 23
3.3 Duomenų transformacijos 25
3.3.1 Vertinimo duomenų apdorojimas 26
3.3.2 Recenzavimo duomenų apdorojimas 26
3.3.3 Reitingavimo duomenų apdorojimas 26
3.3.4 Straipsnių skaitymo duomenų apdorojimas 27
3.3.5 Straipsnių rašymo duomenų apdorojimas 27
3.4 Statistikos modulio architektūra 28
3.4.1 Komponento „KMeans“ detalizavimas 28
3.4.2 Komponento „DotNetCharting“ detalizavimas 30
3.4.3 Komponento „PL“ detalizavimas 31
3.4.4 Komponento „DAL“ detalizavimas 33
3.4.5 Komponento „Procedures” detalizavimas 33
4 STATISTIKOS MODULIO REALIZACIJA 34
4.1 Statistikos modulio veikimas 34
4.1.1 Lankytojams prieinamos funkcijos 34
4.1.2 Registruotiems vartotojams prieinamos funkcijos 39
4.1.3 Duomenų transformacijų veikimas 44
4.2 Įgyvendinimo priemonių parinkimas 46
4.3 Statistikos modulio realizacijos apibendrinimas 46
5 STATISTIKOS MODULIO EKSPERIMENTINIS TYRIMAS IR VERTINIMAS 46
5.1 Duomenų gavybos ir atvaizdavimo savybių palyginimas 46
5.2 Duomenų grupavimo eksperimentai 48
6 IŠVADOS 51
7 LITERATŪROS ŠALTINIŲ SĄRAŠAS 52
8 TERMINŲ IR SANTRUMPŲ ŽODYNĖLIS 54
9 PRIEDAI 55

9.1 Priedas Nr.1 Serveryje saugomos procedūros 55
9.2 Priedas Nr.2 Portalo DB schemos fragmentai 57
9.3 Priedas Nr.3 DB pagrindinių lentelių fragmentai 58
9.4 Priedas Nr.4 Straipsnis konferencijos „Informacinė visuomenė ir universitetinės studijos 2007“ leidinyje 67

Portal Statistics Module Based on Clustering

SUMMARY

Presented data mining methods and clustering usage in current statistical systems and created statistics module prototype for IT knowledge portal for data storage, analysis and visualization. Suggested module prototype can be considered as Data Mining system, which helps to select information from vast amount data and visualize it. The part of prototype is periodically performed data discretizations for the purpose to unweight database performance. Statistical data accessed in portal can be clustered. Clustered information represented graphically may serve for interpreting information when trends may be noticed. One of the best known data clustering methods – parallel k-means method – is adapted for similar data clusters separation. The statistics module enables user to analyze portal performance trends and information that can advise administrator for altering system settings.

1 ĮVADAS

Kompiuterizacijos pažanga ir duomenų kaupimo mastai tiesiog užtvindė mus informacija [1]. Todėl atsirado poreikis pasitelkti į pagalbą naujų technologijų ir įrankių, kurie padėtų transformuoti duomenis į naudingas žinias. Potencialiai svarbi informacija tarp gausių duomenų kartais lieka nepastebėta. Naudingoms žinioms aptikti pasitarnauja duomenų gavybos (DG) įrankiai, kurie gali parodyti duomenų tendencijas ar net pateikti ateities prognozes [2].

Informacinių technologijų (IT) žinių portalo paskirtis – kaupti mokslinius IT srities straipsnius dalyvaujant vartotojams-autoriams, tuo pačiu straipsnius bei autorius reitinguoti, vertinti, taip skatinant portalo plėtrą. Portale numatytas vartotojų skaidymas į lygius pagal jų reitingus. Aukštesnio lygio vartotojai atlieka recenzentų vaidmenį. Jei straipsnių labai daug, o recenzentų mažai, tokiu atveju recenzentui tenka didelis krūvis. Taip pat jei veiksmai, kurie įtakoja reitingo reikšmę, labai intensyvūs, o už juos suteikiama didelė vertė, tai reitingo reikšmė gali išaugti per greitai. Todėl reikalingi tokie skaičiavimai, kurie administratoriui vaidintų patariamąjį vaidmenį, t.y. padėtų valdyti barjerų ribines reikšmes patekimui į kiekvieną lygį, reitingų sudėties elementų ribines reikšmes ir kt. Taip pat yra reikalavimas suteikti lankytojams galimybę palyginti veiksmų aktyvumą portale, vartotojų apkrautumą ir pan.

Darbo tikslas – sukurti IT žinių portalo IT-EUROPE statistikos prototipą duomenų saugojimui, analizei ir peržiūrai atlikti. Portale kaupiami gausūs duomenys, todėl naudinga pateikti jų statistiką, turimus duomenis apie portalo autorius ir straipsnius suskirstyti taip, kad panašiausi objektai būtų išsidėstę arčiausiai vienas kito suskirstyti duomenis grupėmis. Pagal grupuotą informaciją paprastas vartotojas galėtų stebėti portalo veiklos mastus, analizuoti informaciją ir bandyti atrasti grupių skiriamuosius bruožus, o administratorius pagal tai galėtų priimti sprendimus dėl sisteminių apribojimų keitimo. Portalo statistikos naudotojas pirmiausiai nori gauti kuo daugiau informatyvumo, todėl svarbu informaciją pateikti suprantamai, lanksčiai (procentiniai palyginimai, įvairūs kiekiai), leisti peržiūrinėti objektų (straipsnių, autorių) veiksmų istoriją bei jų kitimą laike bei leisti pasirinkimus (laiko intervalo, grupių kiekio).

Paplitusios duomenų analizės ir statistikos sistemos yra gana išvystytos ir naudoja pažangiausius metodus duomenų gavybai, analizei ir prognozavimui [13], [14], [15]. Tačiau šios sistemos dažniausiai taikomos itin svarbių duomenų (medicinos, akcijos rinkos, geografinių) analizei ar prognozavimui ir daugumos jų neina naudoti internetinėje prieigoje. Minėtos sistemos yra didelės ir reikalaujančios nemažai resursų, o mūsų problemai spręsti reikalinga sistema, veikianti internetinėje aplinkoje ir skirta darbui su ne kritiškai svarbiais duomenis.

Darbo tyrimo sritis – statistikos ir duomenų analizės sistemų, naudojančių duomenų gavybos metodus, modelių sudarymas bei internetinių svetainių veiklos statistikos atvaizdavimo ir pateikimo būdai.
Tyrimo objektas – žinių portalo statistikos modulis, diskretizuojantis duomenis, pateikiantis statistiką ir leidžiantis analizuoti duomenis juos grupuojant ir atvaizduojant.

Analizės tikslą nusako tolesni punktai:

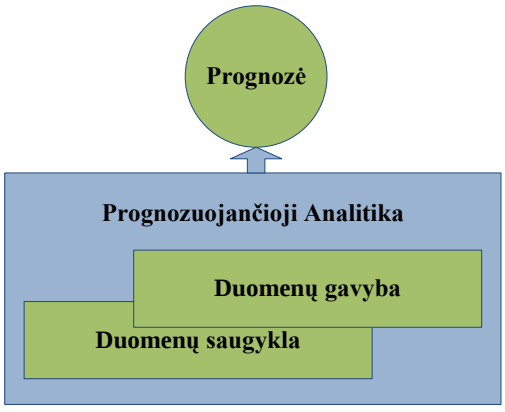
- Duomenų grupavimo apibrėžimas ir metodų analizė;
- Lygiagrečiojo k-vidurkių grupavimo metodo nagrinėjimas;
- Duomenų grupavimo taikymas statistikos sistemoms;
- Statistikos sistemų savybių palyginimas;

Paanalizavus DG metodus buvo pasirinktas grupavimas ir vienas žinomiausių ir net didelėse sistemose [13], [14] dažniausiai naudojamų metodų – lygiagretusis k-vidurkių grupavimo metodas. Buvo išnagrinėtos kelios statistikos sistemos ir jų savybės. Pastebėta, kad jos naudoja platų spektrą naujoviškų ir pažangių technologijų. Kuriama statistikos prototipas pagal grupavimo metodą ir teikiamus vartotojo pasirinkimus grupavimui artimiausias Mineset sistemai, bet orientuotas internetiniam naudojimui.

Didelio kiekio pavienių duomenų išrinkimas statistiniams skaičiavimams gali apsunkinti duomenų bazę, todėl periodiškais laiko momentais duomenis siūloma diskretizuoti, nes realaus laiko sistemos dėl didelio apkrautumo reikalauja daug resursų ir pajėgumo. Transformacijų metu portalo sukaupti duomenys apibendrinami, diskretizuojami ir saugomi statistikos reikmėms.

Statistikos modulio koncepcija aprašoma pateikiant sistemos konceptualiąją schemą, fragmentus iš PĮ specifikacijos, kur aprašomi panaudojimo atvejai, funkcionalumo pjūviai, klasių modelis, detaliai aprašomos šiame sprendime svarbios duomenų transformacijos.

Sukurtas IT žinių portalo statistikos modulio prototipas duomenų saugojimui, analizei ir peržiūrai atlikti gali būti pavadintas DG sistema. Ji padeda atrinkti duomenis iš didelio jų kiekio ir suprantamai pateikti. Duomenims atrinkti ir panašių objektų grupėms išskirti pakoreguotas (pradinių objektų parinkimas) ir pritaikytas vienas iš žinomiausių duomenų grupavimo metodų – lygiagretusis k-vidurkių metodas. Statistikos modulis leidžia peržiūrinėti objektų (straipsnių, autorių) veiksmų ar veiksmų su objektais statistiką laike, grupių išskyrimui gelbsti naudojamas greitas grupavimo metodas [12]. Kuriamas produktas padeda spręsti didelio kiekio informacijos tinkamą atrinkimą bei suprantamą pateikimą peržiūrai ir analizei. Grafiniam atvaizdavimui panaudotas „*DotNetCharting*“ komponentas [19].



1 pav. *Prognozuojančiosios analitikos procesas*

DG ateitis slypi prognozuojančioje analitikoje [2] (Predictive Analytics), kuri sparčiai vystoma ir paklausi dabartinėje rinkoje medicinos, ekonominiams ir kt. procesams prognozuoti. Duomenų saugykla yra architektūra duomenų valdymui. DG – procesas naudojantis duomenų apibendrinimą ir statistikos principus.

Prognozuojančioji analitika (***1 pav.***) apjungia šias dvi technologijas programoje, eksploatuojančioje didelius duomenų kiekius ir teikiančia įvairių sričių prognozes. Todėl parinkus prognozuojančiojo modelio [14] sudarymo metodiką, statistikos modulį galima išplėsti iki prognozuojančiosios sistemos.

2 STATISTIKOS MODELIŲ IR METODŲ ANALIZĖ

2.1 Duomenų gavyba ir grupavimas

Duomenų gavyba

Duomenų kaupimas jų neapdorojus neturi prasmės, todėl duomenų analizė buvo atliekama visais laikais. Visų rūšių duomenų tvarkymas yra pirmas žingsnis norint palengvinti duomenų analizę, apdoroti didelį informacijos srautą. Viena iš informacijos analizės sričių – DG.

„DG yra prasmingų dėsningumų, modelių ir tendencijų radimo procesas dideliuose informacijos kiekiuose, naudojant modelių atpažinimo statistinius bei matematinius metodus“ [3]. Tai procesas didelių kiekių duomenyse automatiškai aptikti šablonus naudojant priemones: klasifikaciją, asociacijos taisyklių išgavimą, grupavimą [4]. DG – tai numatomos, anksčiau nežinomos ir potencialiai naudingos informacijos iš duomenų netrivialus išgavimas [5].

Duomenų gavyba yra labai plati sritis, todėl yra ir daug jai skirtų metodų, algoritmų bei taikomųjų sistemų.

- **Asociacijų paieška –** tai dėsningumų analizė tam tikrose reiškinių ar daiktų grupėse. Pavyzdys – pirkimo analizė. Nagrinėjama, kokios prekės perkamos kartu, kokia tikimybė, kad bus būtent toks derinys ir t.t. Tokio tipo uždaviniai gali būti sprendžiami rengiant reklamos kampanijas, kuriant nuolaidų sistemas.
- **Eiliškumo analizė –** tai dėsningumų paieška atsižvelgiant į laiką. Šiuo atveju svarbu ne tik tai, kokiomis paslaugomis naudojasi klientas, bet ir kokia eilės tvarka. Šis metodas padeda efektyviau teikti paslaugas.
- **Grupavimas (klasterizavimas)** dažnai būna vienas pirmųjų duomenų gavybos žingsnių. Tai visos duomenų aibės suskaidymas į poaibius pagal skiriamuosius bruožus. Tai ir rinkos ar klientų segmentavimas, ir nekilnojamojo turto grupavimas pagal būdingus duomenis, ir daugelis kitų uždavinių.
- **Klasifikavimas** dažnai atliekamas po grupavimo. Kai nagrinėjama aibė jau padalyta į pogrupius, dažnai kyla klausimas, kam priskirti naujus elementus. Grupavimu anksčiau neįvardyti poaibiai išskiriami iš duomenų visumos, o klasifikuojant sprendžiama, kaip sudėti elementus į žinomas grupes.
- Po klasifikavimo atliekamas **įvertinimas** . Pavyzdžiui, finansinė institucija ne tik nori žinoti savo išduotų paskolų apibūdinimą "gera – bloga", bet ir jų įvertinimą.

- **Prognozavimas** taip pat labai svarbus duomenų gavybai. Atsižvelgiant į turimus duomenis bei pastebėtas tendencijas, bandomi prognozuoti ateities įvykiai.

Dažnai uždaviniui išspręsti taikomi keli metodai iš eilės ar net sudėtingi jų deriniai. Nė vienas jų nėra universalus ar nepriekaištingas. Vienų trūkumas – sudėtingumas, kitų – didelė modelių apimtis, daug sugaištama laiko. Bet visų metodų pagrindinė užduotis – atrasti duomenų šablonus, kad duomenis būtų galima saugoti duomenų bazėje (DB) [6].

Grupavimas

Duomenų grupavimas – tai objektų klasifikavimas į skirtingas grupes, tiksliau tariant, duomenų dalinimas į pogrupius, kad kiekviename jų duomenys turėtų bendrų bruožų – dažniausiai tai artimumas pagal kažkokį numatytą atstumo matą. Duomenų grupavimas yra dažnas metodas statistiniam duomenų tyrimui. Metodas naudojamas tokiose srityse kaip save mokančios sistemos, DG, šablonų atpažinimas, vaizdų analizė, bioinformatika ir genų inžinerija, statistika [7].

Šiai sričiai aktualūs pagrindiniai grupavimo metodų tipai yra du: hierarchinis ir padalijimo.

Hierarchiniai metodai

Grupę formuoja hierarchiškai, t.y. kiekvienas grupės viršūnė turi vaikinę grupę. Grupės apjungiamos ir skaidomos taip sudarant hierarchinę struktūrą. Priskiriami klasikiniai SLINK, COBWEB algoritmai bei naujesni CURE ir CHAMELEON algoritmai.

Išskiriami tokie tipai:

- Sujungimo (Agglomerative) algoritmai
- Išskaidymo (Divisive) algoritmai

Padalijimo metodai

Duomenys dalinami į kelis pogrupius. Visų pogrupių patikrinti neįmanoma, todėl naudojama iteracinė optimizacija, kuri palaipsniui gerina grupes.

Išskiriami tipai:

- Kelties (Relocation) algoritmai
- Tikimybinio grupavimo algoritmai: EM (BIRTCH), SNOB, AUTOCLASS, MCLUST;
- Artimiausių kaimynų grupavimo algoritmai;
- K-vidurinių taškų (K-medoids) grupavimo algoritmai: PAM, CLARA, CLARANS;
- K-vidurkių (K-means) grupavimo algoritmai;
- Tankumo algoritmai: DBSCAN, OPTICS, DBCLASD, DENCLUE;

Kiti grupavimo metodai

Yra nemažai naujoviškų metodų: taisyklėmis paremto grupavimo, grupavimo algoritmai save mokančioms sistemoms (Machine learning), gradiantinio kilimo ir dirbtinių neuroninių tinklų (Gradient

Descent and Artificial Neural Networks), prižiūrimojo išmokimo (Supervised Learning), evoliuciniai metodai. Pastarieji naudojami ne tik duomenų gavyboje, bet ir dirbtinio intelekto srityje [8].

2.2 Euklido kvadratinės paklaidos matas

Duomenų grupavime svarbi yra jų panašumo/nepanašumo sąvoka, nes DG technologijos paremtos objektų panašumo matais. Nepanašumas vadinamas atstumu ir matuojamas tam tikrais matais. Dažniausiai sutinkama dalijimo į grupes strategija yra kvadratinės paklaidos kriterijus. Padalijimo metodai dažniausiai išskiria grupę optimizuojant kriterijaus funkciją. Tikslas – fiksuotam kiekiui grupių minimizuoti kvadratinę paklaidą.

Žinomiausias atstumo matas grupavime yra Euklido atstumas. Jis remiasi kvadratinių paklaidų mato principu, kai kiekviena grupė atstovaujama vieno objekto, o kiti priskiriami grupėms pagal panašumą (kuo mažesnis atstumas) [9].

Euklido atstumo skaičiavimo formulė:

$d_E(x_i, x_j) = \sqrt{\sum_{k=1}^{m}(x_{ik} - x_{jk})^2}$, Kur $x_i, x_j \in X$, $(i,j = 1, \ldots, n)$, $x_i = \{x_{i1}, x_{i2}, \ldots, x_{im}\}$ – duomenų vektorius, m – dimensijų skaičius, n – vektorių skaičius;

Euklido atstumas yra atskiras Minkovskio mato atvejis, kai parametras $p=2$.

Minkovskio atstumo apibendrinta formulė:

$$d_M(x_i, x_j) = \left(\sum_{k=1}^{m} \left| x_{ik} - x_{jk} \right|^p \right)^{\frac{1}{p}}$$

Jei parametras lygus $p=1$, turime taip pat gerai žinomą Manheteno atstumą.

2.3 K-vidurkių grupavimo metodas

K-vidurkių algoritmas – vienas iš paprasčiausių, žinomiausių ir dažniausiai naudojamų algoritmų duomenų grupavimui. Jis paremtas kvadratinių paklaidų kriterijumi. Šis algoritmas pateikia efektyvius grupavimo rezultatus daugelyje praktinių taikymo sričių. Šis algoritmas taikomas daugelyje sričių: kalbos atpažinimui, genomų, bioinformatikos duomenų analizei, geografinių informacinių sistemų duomenis [10].

K-vidurkių algoritmo veikimo stadijos:

1. *Inizializacija* - atsitiktinai parenkami *k* objektų iš duomenų rinkinio, taip kiekvienas objektas atstovauja pradinės duomenų grupės vidurkiui (centrui).
2. *Atstumų skaičiavimas* - likę objektai priskiriami grupėms, kurioms yra tinkamiausi pagal Euklido atstumą.
3. *Grupių centrų perskaičiavimas* - grupių centrai, t.y. grupei priskirtų objektų vidurkis.
4. *Konvergavimo sąlyga* – kartojami žingsniai 2 ir 3, kol algoritmas konverguoja.

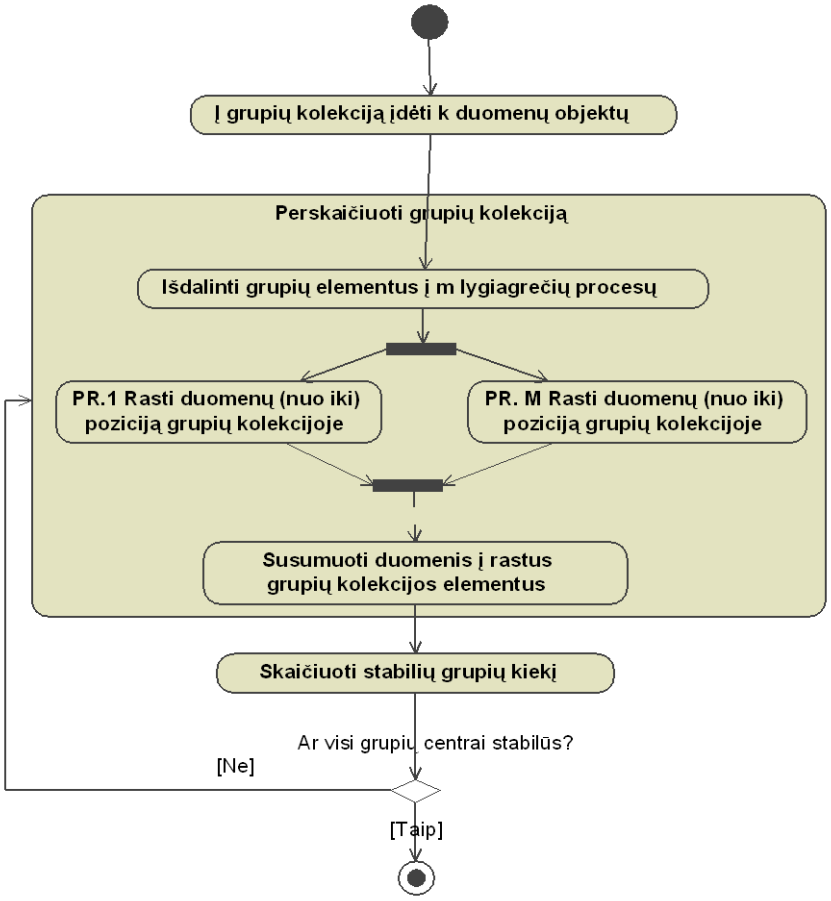
K-vidurkių algoritmo trūkumai:

- Sudėtingumas stipriai veikiamas duomenų objektų kiekio, nes reikia perskaičiuoti atstumus tikrinant visus objektus pakartotinai ir iteratyviai juos pergrupuoti. Kai duomenų kiekis labai didelis, algoritmas užtrunka ilgiau laiko skaičiavimams, tai didina skaičiavimo sudėtingumą ir mažina skaičiavimo greitį duomenų rinkiniuose.
- Pagrindinė algoritmo problema, kad jis jautrus pradiniam grupių skaičiui. Jei šis parenkamas netinkamai, kriterijaus funkcija gali konverguoti į lokalų minimumą. Algoritmui reikėtų mažiau iteracijų, jei pradiniai centrai būtų parenkami pagal geresnę strategiją.
- Naudoja tik skaitines reikšmes.

Lygiagretusis k-vidurkių algoritmas naudoja tą patį metodą atstumams tarp objektų skaičiuoti, tik į pagalbą pasitelkiama *m* lygiagrečių procesų, kuriems išdalijami duomenys. Kiekvienam lygiagrečiam procesui priskiriama po duomenų bloką ir duomenys juose grupuojami lygiagrečiai. Visi skaičiavimų rezultatai surenkami kartu. Lygiagretumas ypač pagerina neefektyviausią k-vidurkių algoritmo iteratyvaus pergrupavimo dalį [11].

Lygiagrečiojo algoritmo veiksmai (***2 pav.***):

1. Priskirti pradinius duomenis k grupių kiekiui.
2. Išskaidyti duomenis į m lygiagrečių procesų, skaičiuoti atstumus tarp kiekvieno duomenų objekto ir k grupių centrų, priskiriant objektą artimiausiai grupei.
3. Surinkti reikšmes iš *m* procesų. Surinkta informacija su skaitinių duomenų sumomis ir duomenų objektų kiekiais grupėse naudojama skaičiuoti grupių centrams. Tada informacija surenkama tai pačiai grupei kiekviename lygiagrečiame procese ir perskaičiuojami nauji centrai kiekvienai grupei.
4. Atnaujinami visų *k* grupių centrai kiekviename procese.
5. Kartoti ankstesnius veiksmus iki konvergavimo



2 pav. *Lygiagrečiojo k-vidurkių algoritmo veiklos diagrama*

Paprastojo k-vidurkių algoritmo skaičiavimo laikas:

$$T^{comp} \approx (3nkd) \cdot \tau \cdot T^{flop}$$

Paprastojo algoritmo skaičiavimo sudėtingumas – $O(n\,k\,\tau)$

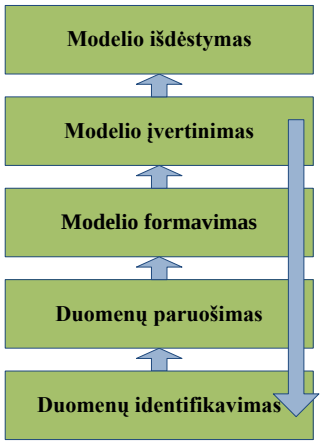
Lyginant su paprastuoju, lygiagretaus algoritmo skaičiavimo laikas sumažinamas m kartų, t.y. kiek naudojamų lygiagrečių procesų [12]:

$T_m \approx T_m^{comp} \approx \frac{(3nkd) \cdot \tau \cdot T^{flop}}{m}$, kur T_m – skaičiavimo sudėtingumas; T^{comp} – skaičiavimo laikas; n – duomenų objektų skaičius; k – duomenų grupių kiekis; d – duomenų vektorių dimensija; τ – iteracijų skaičius; m – lygiagrečių procesų skaičius, duomenų blokų skaičius; T^{flop} – slenkančio taško operacijos (sudėtis, daugyba, palyginimas) laikas;

2.4 Duomenų grupavimo taikymas statistikos sistemoms

2.4.1 Mineset (SGI) – duomenų gavybos ir analizės sistema

MineSet [13] įrankių rinkinys leidžia išgauti, analizuoti ir grafiškai atvaizduoti duomenis, kad juos būtų galima tirti ir suprasti. DG įrankis automatiškai suranda šablonus, atranda tendenciją ir sukuria modelį, kurį galima peržiūrėti atvaizdavimo įrankiais.



3 pav. *DG procesas*

MineSet naudojamas DG procesas (***3 pav.***):

1. Duomenų identifikavimas

Pirmiausiai reikia žinoti, kokių duomenų reikia problemos sprendimui. Duomenys gali būti skirtingose vietose, dažnai – keliose tarpusavyje nesuderinamose DB. Nuo egzistuojančių duomenų priklauso naujų duomenų surinkimo forma. Mineset palaiko sąsajas prie keleto komercinių DB (Oracle, Informix, SQL), ODBC bei duomenų nuskaitymą iš skirtingų failų formatų (Excel, SPSS, kintamųjų, MineSet binarinių failų).

2. Duomenų paruošimas

Duomenys turi būti modifikuoti prieš juos įkeliant į MineSet. Jie gali būti nesuderinami su programa, klaidingi, pasenę ar nepilni, todėl duomenys transformuojami: pridedant stulpelį jį matematiškai išskaičiuojant, pašalinant perteklinį stulpelį, dikretizuojant tolydžius duomenis, sugrupuojant duomenis ar suskaičiuojant papildomas reikšmes, atrenkant atsitiktinius duomenis, klasifikuojant.

3. Modelio formavimas

Tai yra žinių gavybos esmė, automatiškai atliekama analitinių DG algoritmų.

4. Modelio įvertinimas

MineSet naudojami keturi modelio įvertinimo metodai: klaidų nustatymas, sutrikimų matrica, kilimo kreivė, investicijų grąžos kreivės (ROI).

5. Modelio išdėstymas

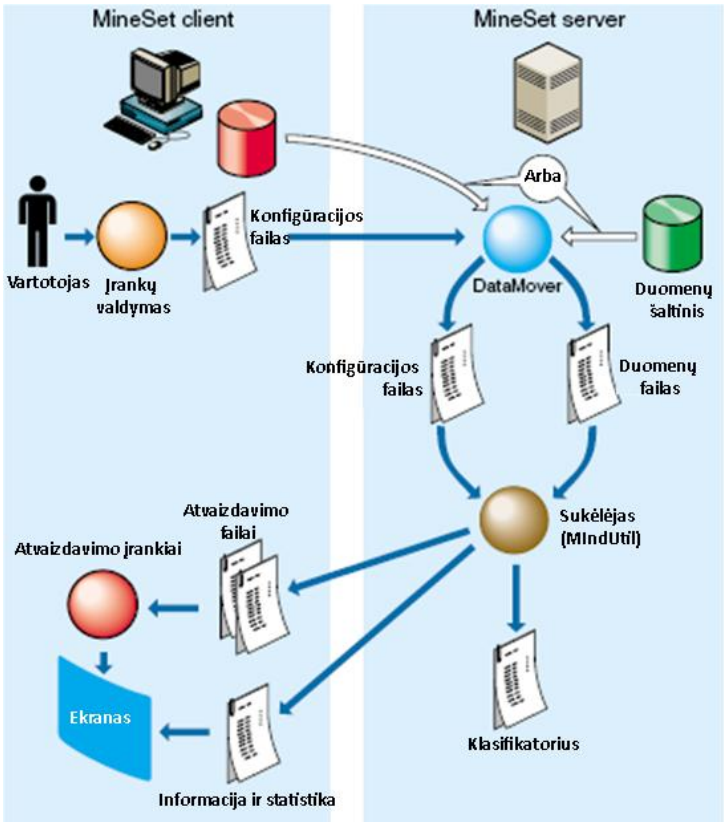
Proceso taikymas specialiai DB

MineSet naudojami analitiniai duomenų gavybos algoritmai

Šie algoritmai automatiškai sukuria modelį iš duomenų. Dažniausiai naudojamos dvi modeliavimo algoritmų šeimos:

1. Prižiūrimasis modeliavimas

Prižiūrimosios užduotys yra prognozuojamojo modeliavimo (***4 pav.***) užduotys, kurių tikslas yra numatyti vieno stulpelio reikšmę pagal kitų stulpelių reikšmes. Šiame modeliavime yra savybė, vadinama etikete (label). Tai – stulpelis, pagal kurį ketinama prognozuoti.



3 pav. *DG prižiūrimojo modeliavimo schema*

Užšifravus ryšį tarp etiketės ir kitų atributų, modelis gali prognozuoti apie naujus, etikete nepažymėtus duomenis. Dažniausios šio modelio užduotys yra klasifikacija ir regresija. Jei etiketė yra diskreti (turinti fiksuotą rinkinį reikšmių), užduotis vadinama klasifikacija. Jei tolydi – regresija.

- Regresija

  Regresijos pavyzdys – numatyti atlyginimą ar akcijos kainą yra regresija, o prognozuoti, ar atlyginimas yra numatytose ribose, ar akcijų kaina pakils/nukris, - klasifikacijos užduotis.

- Klasifikacija

Įrašai sudalinami į iš anksto numatytas grupes. Pavyzdžiui, paprasta klasifikacija gali grupuoti sąskaitų įrašus į dvi grupes: kurias reikia apmokėti per 60 dienų ir kurias daugiau nei per 60 dienų. Taip pat galima klasifikuoti pagal lytį, pajamas ir pan. Klasifikacija taip pat gali numatyti tikimybę, kad etiketės atributas įgaus tam tikrą reikšmę tam tikram įrašui. (pvz.: tikimybė, kad klientas apmokės sąskaitą per 60 dienų, kai pateikiamos kitos kliento įrašų reikšmės).

Klasifikatorius yra modelis numatantis vieną atributą iš duomenų rinkinio, kai pateikiami kiti atributai. MineSet gali išgauti klasifikatorių automatiškai iš duomenų pogrupio, vadinamo rengiamuoju rinkiniu (training set). Taip pat galima modelį ir klasifikatoriaus veikimą atvaizduoti trimačiame grafike.

Sukėlėjas (inducer) yra algoritmas, kuriantis prognozuojantį modelį iš rengiamojo rinkinio, sudaryto iš įrašų su etiketėmis. Rengiamasis rinkinys yra duomenų rinkinio, naudojamo sukėlėjo modeliui konstruoti, dalis. Modelio struktūra gali būti vizualizuota ar panaudota klasifikuoti nesužymėtiems įrašams. MineSet sukėlėjai vykdomi MineSet serveryje, nes tai procesorių apkraunantys procesai.

Sukėlėjams reikalingas rengiamasis rinkinys – lentelė su atributais ar charakteristikos, iš kurių viena pažymėta etikete. Modelis gali prognozuoti šią žymę naujiems įrašams. Naujieji įrašai turi būti lentelėje su tais pačiais kaip modelyje atributais, jų pavadinimais ir tipais, bet lentelėje etiketė nereikalinga. Prognozavimo metu ji ignoruojama.

MineSet turi sukėlėjus keturiems klasifikacijos modeliams:

- Sprendimų medžių
- Pasirinkimų medžių
- Pagrindimo ir sprendimų lentelių klasifikatoriams

2. Neprižiūrimasis modeliavimas

Neprižiūrimosios užduotys yra aprašomojo modeliavimo užduotys, kurių tikslas – atrasti šablonus ir duomenų dalis pagal elgsenos panašumą, tai – aprašymo užduotis, o ne prognozavimo. MineSet teikia du šio tipo modeliavimo metodus: asociacijų ir grupavimo.

❖ Asociacijos

Generuoti asociacijas – tai užduotis nustatyti taisykles sąsajos tarp duomenų atributų A ir B. Asociacijos dažnai naudojamos rasti gimiringoms grupėms rinkos krepšelių analizei, prognozuoti su kuriomis prekėmis kitos perkamos kartu dažniausiai.

❖ Grupavimas

Grupavimo algoritmai skaido duomenis į įrašų grupes pagal panašias charakteristikas. Pvz.: tam tikros amžiaus grupės žmonės, dirbantys tam tikroje rinkos sferoje, turintys tam tikrą vaikų skaičių, uždirba tam tikrą atlyginimą; tuomet į šią žmonių kategoriją gali būti geriau orientuojami gyvybės draudimų paketai ir pan.

MineSet įrankiai duomenų išgavimui

- Sprendimo medžio sukėlėjas ir vizualizatorius – išgauna klasifikatorių ir atvaizduoja pagal jį gautą sprendimų medį;
- Pasirinkimo medžio sukėlėjas ir vizualizatorius – išgauna klasifikatorių ir atvaizduoja medį, taip pat sukuria alternatyvius pasirinkimus ir klasifikacijos metu jiems suskaičiuoja vidurkius;
- Pagrindimo sukėlėjas ir vizualizatorius – sukuria savo klasifikatorių ir atvaizduoja duomenų požymius;
- Sprendimų lentelių sukėlėjas ir vizualizatorius – sukuria hierarchinį atvaizdavimą, kiekvienoje dimensijoje rodantį poras;
- Grupavimas – grupuoja duomenis pagal charakteristikų panašumą;
- Regresijos medis – sukuria regresorių, kuris prognozuoja realią reikšmę, t.y. gauna reikšmių pakopas, o ne tam iš anksto numatytas ribas;

Grupavimas su MineSet

Grupavimo įrankyje yra tokie pasirinkimai:

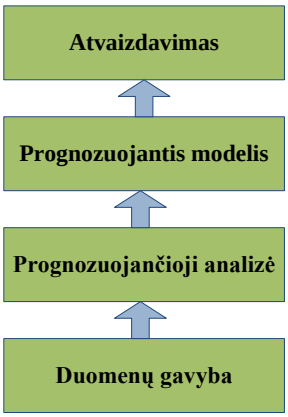
- *Modelio* pasirinkimas (paprastasis k-vidurkių ir iteratyvusis k-vidurkių)
- Paprastajam k- vidurkių metodui galima pasirinkti *grupių kiekį* (numatytoji pradinė reikšmė – 3) ir *maksimalų iteracijų kiekį* (numatytoji reikšmė – 20)
- Iteratyviajam k - vidurkių metodui galima pasirinkti *grupių kiekio intervalą* (numatytasis intervalas: 1-10) ir *pasirinkimo tašką* tarp 0 ir 1, kuris padeda pasirinkti galutinį grupių kiekį (numatytoji reikšmė 0,5); Pagal iteratyvųjį algoritmą pradžioje priskiriama minimalioji grupių kiekio reikšmė ir visi veiksmai kartojami, kol nepasiekiama maksimalioji grupių kiekio reikšmė. Randama grupė su didžiausia dispersija ir dalinama pusiau į dvi grupes, perskaičiuojami grupių centrai ir įrašai, artimesni kitos grupės centrui, perkeliami į artimesniąją grupę
- *Atributo svoris*, kuris suteikia galimybę kiekvienam duomenų rinkinio atributui suteikti svorį.
- *Atstumo matas* – tai būdas, kuriuo matuojami atstumai tarp duomenų. Numatytasis matas – Euklido atstumas. Galimas kitas pasirinkimas – Manheteno matas.
- *Atsitiktinis pradinių reikšmių suteikimas*

MineSet duomenų vizualizavimo įrankiai:

- Žemėlapių atvaizdavimas – duomenys dažnai atvaizduojami geografiniuose žemėlapiuose;
- Išsibarstymo atvaizdavimas – duomenų taškai parodomi 1 – 3 dimensijose;
- Pozavimo (Splat) atvaizdavimas – duomenų skirtumai išskiriami spalviniu debesiu, priklausomai nuo duomenų tankumo;
- Medžių atvaizdavimas – duomenys suvedami į mazgus parodyti hierarchinius išskirstymus
- Diagramos, histogramos ir kt.

2.4.2 STATISTICA Data Miner (StatSoft) – duomenų gavybos, analizės, prognozavimo sistema

STATISTICA Data Miner [14] sistemos paskirtis: atskleisti paslėptas tendencijas, paaiškinti žinomas struktūras, prognozuoti elgseną. Ji apima daug pilnai integruotų pažangių DG ir prognozavimo metodų (grupavimo technologijos, neuroniniai tinklai, save mokančių sistemų architektūros, klasifikacijos/regresijos medžiai ir kt.) bei grafinių vizualizacijos procedūrų.



4 pav. *Prognozuojančios sistemos veiklos etapai*

4 pav. matomas prognozuojančios sistemos veiklos etapai. Pagal tai suskirstyti ir sistemos įrankiai: duomenų klasifikavimui, modeliavimui, prognozavimui ir atvaizdavimui.

Naudojamas platus DG metodų pasirinkimas:

1. Klasifikatorius

Naudojamos technologijos yra duomenų gavybos metodai, daugiklių analizė (Factor Analysis), klasifikavimo medžiai, bendri klasifikavimo ir regresijos medžių modeliai (GTrees), naudojantys CART metodo realizaciją, bendrieji tiesiniai modeliai GLM (General Linear Models) ir GLZ (Generalized Linear Models), bendrieji regresijos modeliai GRM (General Regression Models), bendrieji diskriminanto analizės modeliai GDA (General Discriminant Analysis Models), dalinių kvadratinių paklaidų (PLS), CHAID modeliai, naudingi tiriamajai duomenų analizei ir prognozavimui.

Iš grupavimo metodų šis įrankis naudoja medžių grupavimo, K-vidurkių grupavimo analizę ir praplėstą tikimybės maksimizavimo (EM) metodą su v-fold maišymo ir kryžminio tikrinimo pasirinkimu optimizuotam grupių kiekiui, diskriminanto analizės modelius. Šie grupavimo metodai skirti apdoroti dideliems duomenų rinkiniams juos grupuojant ir suteikiant pagrindą šablonų atpažinimui. Pažangūs EM grupavimo metodai siejami su tikimybinio, statistinio grupavimo galimybėmis.

2. Modeliuotojas

Duomenų gavybos išdėstymo modeliams naudojamas neuroninis tinklas ir ne tokie populiarūs duomenų gavybos metodai: dalinis mažiausių kvadratinių paklaidų metodas – bruožų atrinkimui ir kintamųjų kiekio sumažinimui, natūralioji analizė (survival analysis) – analizuojant detalią informaciją kaip medicininius

tyrimus, struktūrinių lygčių modeliavimo technologija, atsako analizė – sudėtingų lentelių struktūrai analizuoti, daugiklių analizės, daugiamatis matavimas.

3. Prognozuotojas

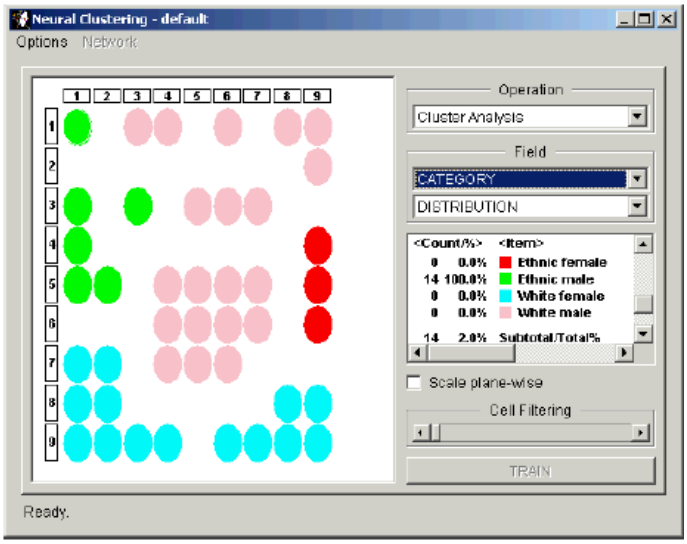
Taikomi ARIMA, eksponentinio glodinimo, Fourier spektrinio skaidymo, regresijos ir polinominių vėlinimų analizė, neuroniniai tinklai. Įrankis gali nagrinėti daugumą duomenų gavybos problemų: klasifikaciją, paslėptos struktūros aptikimas, prognozavimas. Naudoja automatinį pagalbininką su naujausiais neuroninių tinklų architektūromis ir procedūromis, dirbtinio intelekto metodais, gerai optimizuotais algoritmais: daugiasluoksniais suvokiniais, spindulinių tinklų, tikimybinių neuroninių tinklų, save organizuojantys bruožų žemėlapiai, tiesiniai modeliai, grupių tinklai. Galimas kelių modelių rezultatų palyginimas.

2.4.3 StarProbe Data Miner (Rosella) – duomenų gavybos, analizės, prognozavimo sistema

StarProbe Data Miner [15] paskirtis – neuroninis grupavimas, tikimybinis modeliavimas, duomenų analizė.

StarProbe palaiko platų grupavimo algoritmų pasirinkimą ir naujoviškus priemones tikimybiniam modeliavimui:

- Neuroninis grupavimas - naujoviškas įrankis grupavimui naudojantis neuroninius tinklus (***5 pav.***)/ save organizuojančius žemėlapius (SOM);

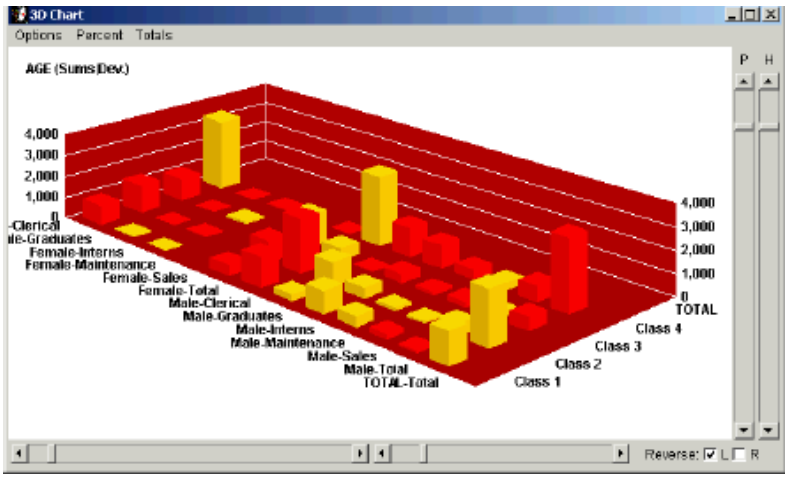


5 pav. *Neuroninio grupavimo atvaizduoti rezultatai*

- Segmentacija – analizė (vizualizacija, grupių profiliavimas, tikimybinis modeliavimas)

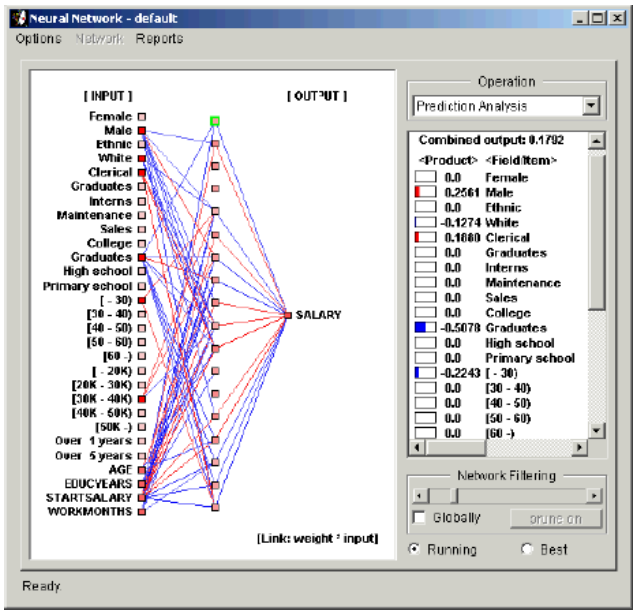
Šiame įrankyje segmetacija naudojama ne tik grupavimui, bet ir statistinių įverčių tikimybiniam modeliavimui. Segmentacijai naudojamas tikimybinis segmentacijos modeliavimas. Tinklas išmokomas grupavimo šablonų ir tada taikomas duomenų segmentacijai ir statistinių reikšmių prognozavimui;

- Karštų taškų analizė – vizualinė tinklų analizė, tinklo įvertinimas, reikšmės numatymas (***6 pav.***);



6 pav. *Karštų taškų analizės atvaizduoti rezultatai*

- Kryžminių lentelių analizė;
- Sprendimų (Crammer) medžiai – struktūra, kur kiekvienas išsišakojimas reiškia vieną ar kitą atsakymą į klausimą. Tokiu būdu sudaromos taisyklės, kurios nagrinėjam duomenų aibę klasifikuoja atsižvelgdamos į elementų savybes. Proceso pradžioje turėta duomenų aibė skaidoma į šakas, kol kiekviena jų tampa homogeniška. Sprendimų medžių privalumas — jų aiškumas, jie suprantami tiek problemą formuluojantiems, tiek bandantiems uždavinį realizuoti, tiek analizuojantiems gautus rezultatus;
- Neuroniniai tinklai - tai netiesiniai modeliai, sėkmingai naudojami klasifikavimo ir prognozavimo uždaviniams spręsti. Jų struktūra primena biologinius neuroninius tinklus. Skirti intuityviam tinklo atvaizdavimas (***7 pav.***). StarProbe grupavimui bei tikimybiniams modeliams kurti naudojami neuroniniai tinklai (save organizuojantys žemėlapiai);



7 pav. *Neuroninio tinklo atvaizdavimas*

- Regresija;
- Taisyklių indukcija – taisyklėmis paremtas tikimybinis modeliavimas ir modelio įvertinimas: grupių profiliavimas ir sprendimų taisyklės
- Vizualizacijos: stulpelinės diagramos, skritulinės, 3D stulpelinės, histogramos, zonų diagramos ir kt.
- Statistika: parametrinė ir neparametrinė statistika, vienpusė ANOVA ir koreliacijos analizė.

2.4.4 Internetinių portalų sprendimai

❖**IT profesionalų portalas** [16]

IT profesionalų portalas – tai naujienų portalas su diskusijomis ir jų statistika. Yra realizuotas PHP kalba.

Šiame analoge noriu apžvelgti pateikiamus gan paprastus duomenų analizės būdus (***8 pav.***):

- Bendroji diskusijų informacija: tema, žinučių kiekis, vidurkis;
- Pateikiami diskusijų temų sąrašai: tema, autorius, peržiūrų kiekis, atsakymų kiekis, paskutinio pranešimo autorius ir data;
- Aktyviausių dalyvių sąrašas: vardas, pranešimų kiekis;
- Aktyviausios, žiūrimiausios temos: tema, atsakymų kiekis, autorius;
- Narių, temų autorių, temų dalyvių populiarumo sąrašai: vardas, pranešimų kiekis, procentinė reikšmė;

Top nariai

Reitingas	Vardas	Pranešimai	%
1	CzV	1349	23,33%
2	/dev/null	835	14,44%
3	NoName	330	5,71%
4	Tereru	260	4,5%
5	KOZERIS	185	3,2%
6	gdrs	146	2,52%
7	Ernetas	124	2,14%
8	arunasroot	117	2,02%
9	hangover	115	1,99%
10	pjanas	113	1,95%

8 pav. *Statistikos pateikimo būdas portale*

❖**Gyvenimo aprašymų sistema „CV.lt“** [17]

Sistemoje pateikiami duomenys anoniminiam vartotojui:

- Užregistruotų CV skaičius
- Naujausių CV skaičius
- Per dieną naujų CV skaičius
- Nuolatinių darbų skaičius, taip pat pateikiamas darbų sąrašas:

Skelbimo numeris, kompanijos pavadinimas, pareigų pavadinimas, miestas, iki kada skelbimas galioja;

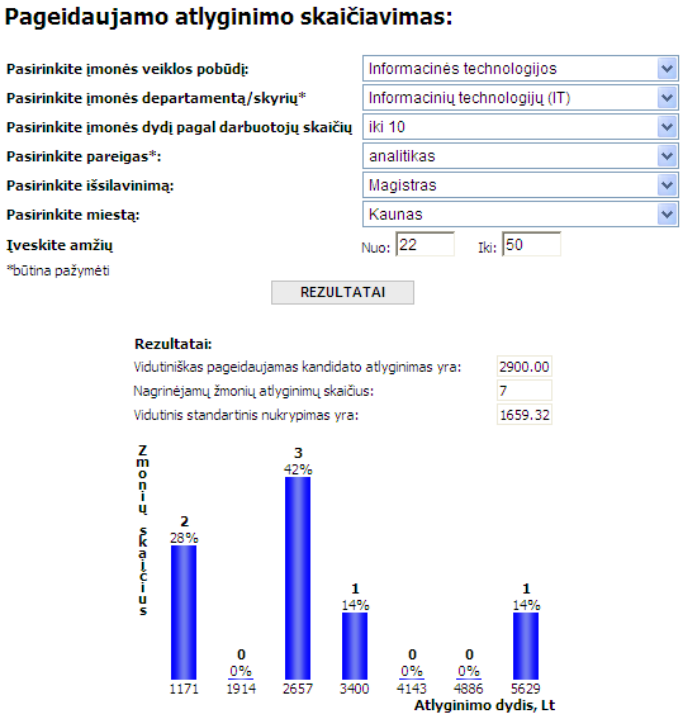
- Papildomų darbų skaičius, taip pat pateikiamas darbų sąrašas:

Skelbimo numeris, kategorija, skelbimo antraštė, paskelbimo data, galiojimo data, kiek kartų skelbimas paržiūrėtas;

- Darbdavių/agentūrų skaičius
- Daugiausiai darbuotojų ieškančių darbdavių penketukas

Sistemos informacija prisijungusiam vartotojui pateikia tokią papildomą informaciją:

- Kiek kartų iš viso buvo prisijungta
- Paskutinė CV taisymo data
- Kiek kartų taisytas CV
- Kiek kartų atsakyta į skelbimus
- Kiek kartų siųsta darbdaviui
- Kiek kartų CV žiūrėjo darbdaviai
- Kiek kartų CV pasirinko darbdaviai



9 pav. *Rezultatų pateikimas modifikuota histograma*

9pav. parodyta modifikuota histograma, kur nežinomu metodu parenkami pradinis ir galinis rėžiai bei žingsnis. Pastebima, kad histogramos stulpelių kiekis priklausomas nuo nagrinėjamų žmonių skaičiaus. Kiekviename stulpelyje matomas patekimo į jį dažnis (žmonių kiekis).

❖ **Klasiokų bendravimo portalas „Klase.lt”** [18]

Anoniminiam lankytojui matoma tokia sistemos statistikos informacija:

- Prisijungusių sistemos narių skaičius
- Tik naršančių narių skaičius
- Bendras narių skaičius sistemoje
- Mokyklų, kuriose užsiregistravo daugiausia klasių, sąrašas:

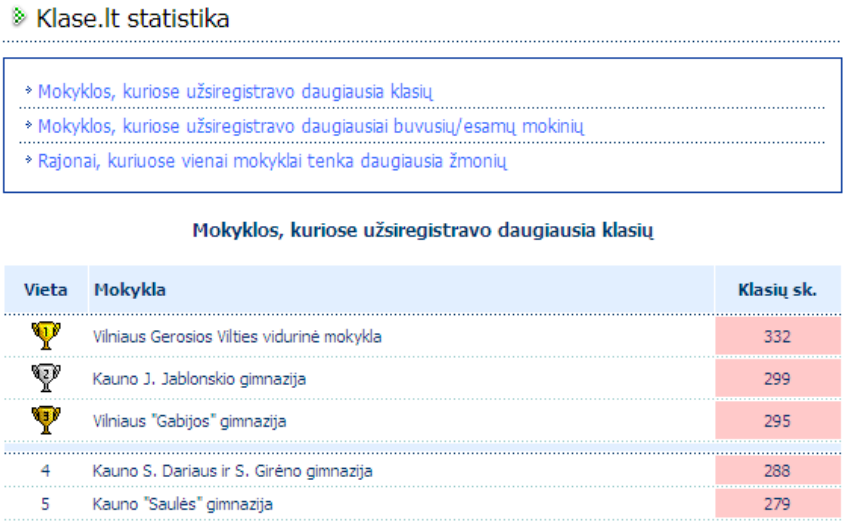
Užimama vieta, mokyklos pavadinimas, užsiregistravusių mokyklos klasių skaičius;

- Mokyklų, kuriose užsiregistravo daugiausiai buvusių/esamų mokinių, sąrašas:

Užimama vieta, mokyklos pavadinimas, užsiregistravusių mokyklos mokinių skaičius;

- Rajonų, kuriuose vienai mokyklai tenka daugiausia žmonių, sąrašas:

Užimama vieta, misto/rajono pavadinimas, vidutinis mokinių kiekis rajono mokykloms (***10 pav.***);



10 pav. *Statistikos pateikimas*

2.5 Siūloma portalo duomenų analizės sistema

Kuriama sistema neprilygs galingiems duomenų gavybos ir analizės įrankiams kaip MineSet, Statistica ar StarProbe, nes ji bus orientuota siauros srities portalui, kur tokių galingų ir daug funkcijų teikiančių įrankių neprireiks. Pakaks pasirinkti vieną efektyvų grupavimo algoritmą. Numatyta pasirinkti k-vidurkių lygiagretųjį grupavimo algoritmą.

Kaip matėme MineSet sistema kaip tik grupavimui naudoja dvi k-vidurkių algoritmo versijas ir draugiškos vartotojo sąsajos pasirinkimų principus. Tai labai geras bruožas.

Statistica sistema naudoja platų gavybos metodų pasirinkimą, o vienas iš grupavimo metodų yra k-vidurkių grupavimo analizė. Prognozavimui šis įrankis naudoja pažangias technologijas.

StartProbe sistema grupavimui naudoja platų spektrą algoritmų, o prognozavimui pasitelkti taip pat naujoviški ir pažangūs metodai.

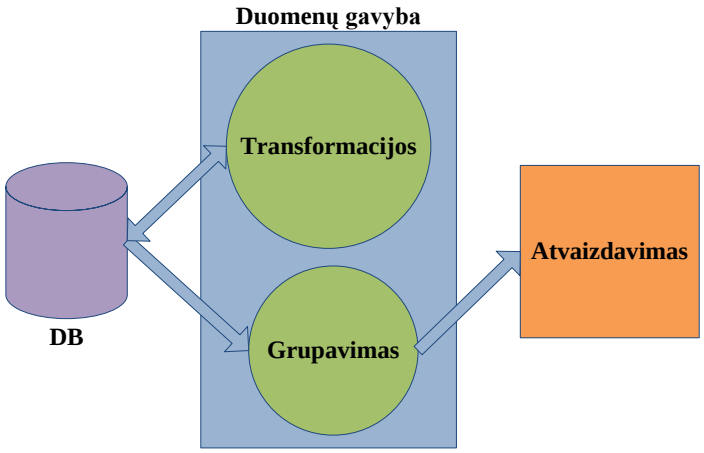
Nagrinėtos sistemos naudoja pažangius atvaizdavimo įrankius, tačiau portalui pakanka tradicinių atvaizdavimo būdų: diagramų histogramų ar paprastų grafikų. Toks sprendimas priimtas, nes nebus analizuojama sudėtinga kelių pjūvių informacija. Visų komponentų kurti iš naujo tikrai neapsimoka, todėl

geriausia panaudoti jau sukurtą ir rinkoje siūlomą grafinio atvaizdavimo komponentą. Vienas iš tokių įrankių – „*DotNetCharting*“ komponentas [19], kuris gali būti naudojamas nemokamai.

Reikalingas duomenų diskretizavimas, todėl svarbu reguliariai saugoti suminius duomenis. Reikia padaryti lanksčią sistemą, kad įvykus klaidai, kai duomenys neišsisaugos ar bus išsaugoti kelis kartus, duomenys nenusigadintų.

3 STATISTIKOS MODULIO KONCEPTUALIOJI SPECIFIKACIJA

Statistikos modulio koncepciją galima pavaizduoti (***11 pav.***) kaip DB pirmoje grandyje, DG priemones, kontaktuojančias su DB, vidurinėje grandyje ir atvaizdavimo priemones paskutinėje grandyje. Duomenų transformacijos vyksta kasdien, o to proceso metu portalo lentelių duomenys diskretizuojami, apibendrinami ir atitinkamais pjūviais saugomi statistikai skirtose lentelėse. Grupavimas („*KMeans*“ komponentas) ir atvaizdavimas („*DotNetCharting*“ komponentas) vyksta vartotojui naršant ir pasirenkant peržiūras, reikalaujančias grupavimo. Rezultate gaunami atvaizduoti duomenys.

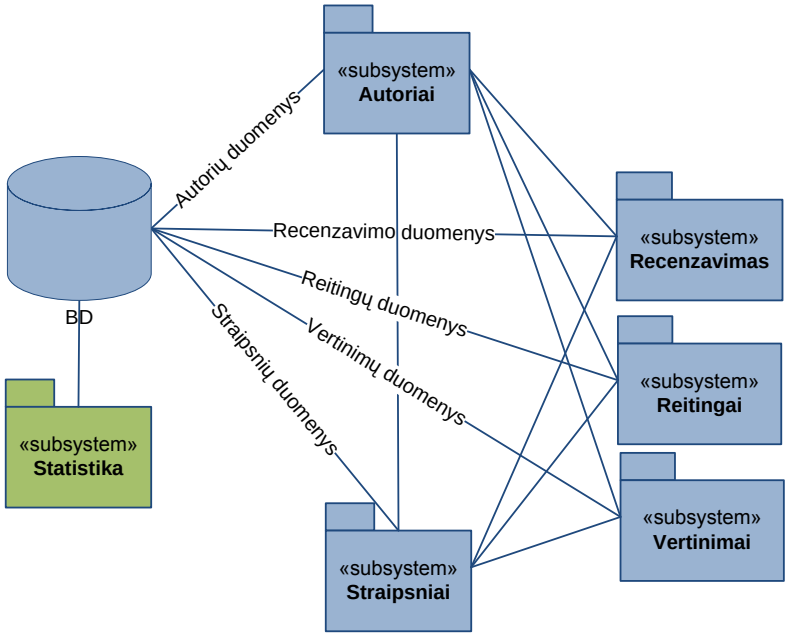


11 pav. *Statistikos modulio koncepcinė schema*

3.1 Funkcionalumas

3.1.1 Veiklos kontekstas

Veiklos kontekstui įtakos turi informacijos srautai ateinantys iš visų portalo posistemių, kurių duomenis statistikos posistemė ima iš saugyklos (***12 pav.***), juos apdoroja, saugo į statistikos lenteles, taip pat juos gali atitinkamai pateikti peržiūrai.



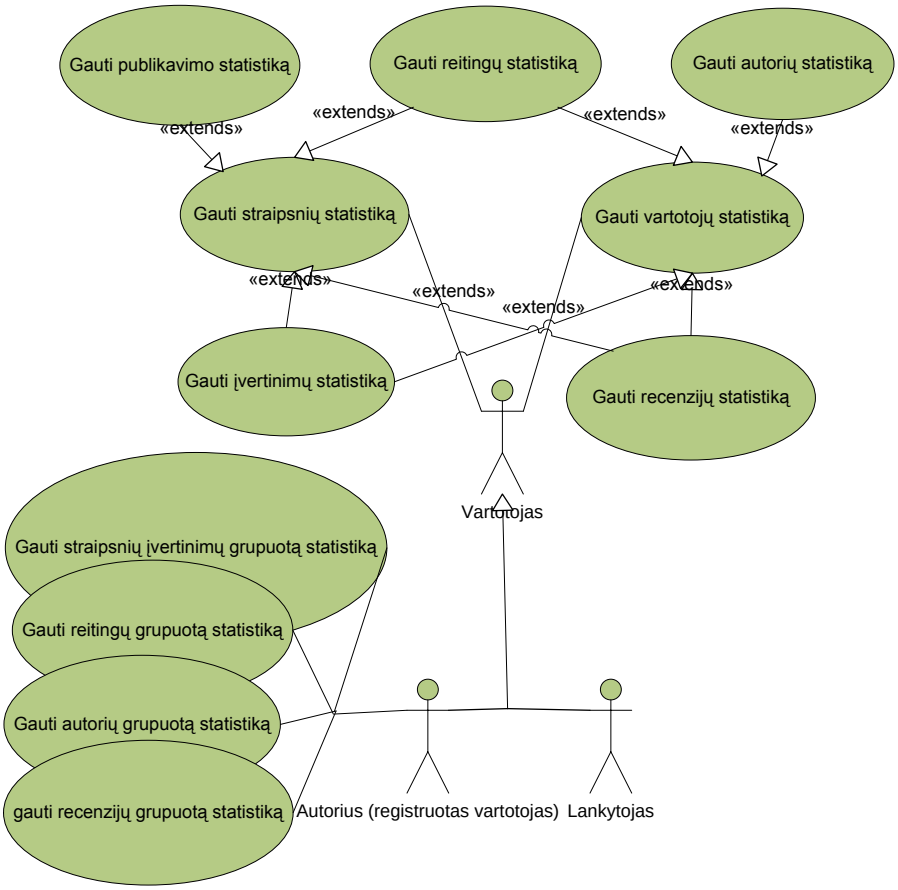
12 pav. *Statistikos posistemio konteksto diagrama*

3.1.2 Sistemos panaudojimo atvejai

Portalo vartotojams prieinamos detalios ir bendros statistikos peržiūros apie portalo objektus (straipsnius ir autorius) bei veiksmus ir savybes (autorių parašytų straipsnių kiekius, autorių parašytų recenzijų kiekius, atliktų vertinimų kiekius, autorių/straipsnių pasiektų reitingų dydžius, straipsnių recenzijų kiekius, straipsnių skaitymų kiekius).

Portale veiks tokie vartotojai (***13 pav.***):

- Lankytojai – nesvarbūs (atsitiktiniai) vartotojai. Jie gaus tik paviršutinišką, bendrą informaciją apie portalo objektus ir įvykius.
- Registruoti vartotojai – pirmaeiliai vartotojai. Jie gali gauti detalesnę statistikos informaciją;
- Administratoriai – svarbiausi vartotojai. Jų nuomonė ir pageidavimai svarbiausi priimant sprendimus portalo veikoje, atsižvelgiant ir panaudojant statistinius duomenis;



13 pav. *Sistemos panaudojimo atvejai*

3.1.3 Funkcionalumo pjūviai

Portalo funkcionalumas logiškai gali būti suskaidomas dviem pjūviais: straipsnio arba autoriaus, o kiekvienas iš jų – dar bendros, vertinimų, recenzijų, reitingų informacijos pjūviais. Tai galima pavaizduoti apibendrinta registruotam/neregistruotam vartotojui teikiamų funkcijų išklotine (***1 lentelė***) apie objektų (straipsniai, autoriai) dalių (bendros, vertinimų, recenzijų, reitingų) informacijos analizę.

1 lentelė. *Funkcionalumo pjūviai*

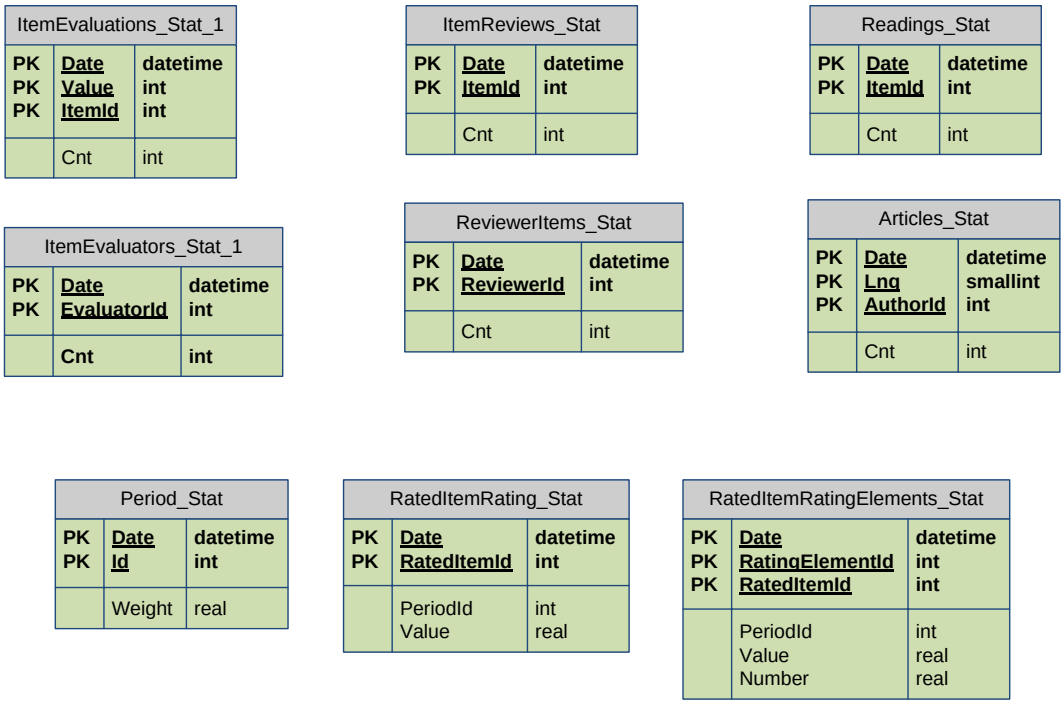
Obj. dalis	**Obj.**	**Straipsniai**		**Autoriai**	
	Vart.	**Neregistruotas**	**Registruotas**	**Neregistruotas**	**Registruotas**
Bendra informacija		Skaitomiausių, naujausių straipsnių populiarumo sąrašai laike; Skaitytų, rašytų straipsnių kiekiai laike; Straipsnių kiekio sistemoje informacija laike;	Autoriaus parašytų straipsnių kiekiai laike; Straipsnių skaitymų kiekiai sugrupuoti bendrame kontekste;	Daugiausiai, artimiausiu metu rašiusių autorių populiarumo sąrašai laike; Autorių kiekio sistemoje informacija laike;	Autoriaus kiekviena kalba rašytų straipsnių kiekiai laike; Autorių parašyti straipsnių kiekiai grupuoti bendrame kontekste; Santykinė informacija autoriui;

Vertinimų informacija	Daugiausiai, geriausiai vertintų straipsnių populiarumo sąrašai laike; Sukaupti straipsnių vertinimai laike; Kasdien gauti straipsnių vertinimai laike;	Autoriaus parašytų straipsnių vertinimai laike;	Daugiausiai vertinimų atlikusių autorių populiarumo sąrašai laike;	Autoriaus atliktų vertinimų kiekiai laike; Autorių atliktų vertinimų kiekiai sugrupuoti bendrame kontekste;
Recenzijų informacija	Labiausiai recenzuojamų, naujausiai recenzuotų straipsnių populiarumo sąrašai laike;	Autoriaus straipsnių recenzentų kiekiai laike; Straipsnių recenzentų kiekiai sugrupuoti bendrame kontekste; Autoriaus straipsnių recenzentų informacija laike;	Daugiausiai recenzuojančių autorių populiarumo sąrašai laike;	Autoriaus recenzuotų straipsnių kiekiai laike; Autorių recenzijų kiekiai sugrupuoti bendrame kontekste; Autoriaus recenzijų informacija laike;
Reitingų informacija	Geriausio reitingo straipsnių populiarumo sąrašai laike;	Autoriaus parašytų straipsnių reitingo vertė laike; Straipsnių pasiektų reitingų periode grupavimas;	Geriausio reitingo autorių populiarumo sąrašai laike;	Autoriaus reitingo vertė laike; Autorių pasiektų reitingų periode grupavimas;

3.2 Duomenų vaizdas

Statistikos posistemis naudojasi visos sistemos duomenų baze. Įrašai apie vykstančius įvykius fiksuojami bendrose portalo DB lentelėse, o statistinei informacijai kaupti iš jų paimami, diskretizuojami ir keliais pjūviais saugomi statistikos lentelėse. Bendras portalo duomenų modelis sudėtingas, todėl jo nepateiksiu, bet ***priede Nr.1*** galima matyti bendros portalo DB fragmentus, kurie artimai susiję su statistikos moduliu.

Statistikos modulio duomenų modelis matomas ***14 pav***.



14 pav. *DB statistikos lentelės*

DB lentelių paaiškinimai pateikti ***2-10 lentelėse***.

2 lentelė. *DB lentelės „Evaluators_Stat“ aprašymas*

Kaupiama tik per nurodytą dieną autorių vertinimo statistika		
Laukas	**Tipas**	**Aprašymas**
Date	*datetime*	Datai kasdien saugoti
Cnt	*int*	Autoriaus vertinimų kiekis per dieną
EvaluatorId	*uniqueidentifier*	Autoriaus ID

3 lentelė. *DB lentelės „ItemEvaluations_Stat“ aprašymas*

Straipsnių gautų vertinimo reikšmių statistika per nurodytą dieną		
Laukas	**Tipas**	**Aprašymas**
Date	*datetime*	Datai kasdien saugoti
Value	*int*	Vertinimo balas
Cnt	*int*	Straipsnių vertinimų kiekis atitinkamu balu per dieną
ItemId	*uniqueidentifier*	Straipsnio ID

4 lentelė. *DB lentelės „ReviewerItems_Stat“ aprašymas*

Autorių recenzijų statistikos lentelė. Saugo autorių parašytų recenzijų kiekius per dieną		
Laukas	**Tipas**	**Aprašymas**
Date	*datetime*	Datai kasdien saugoti
Cnt	*int*	Autoriaus recenzijų kiekis per dieną
ReviewerId	*uniqueidentifier*	Autoriaus ID

5 lentelė. *DB lentelės „ItemReviews_Stat“ aprašymas*

Straipsnių recenzijų statistikos lentelė. Saugo straipsnių gautų recenzijų per dieną		
Laukas	**Tipas**	**Aprašymas**
Date	*datetime*	Datai kasdien saugoti
Cnt	*int*	Recenzijų straipsniui kiekis per dieną
ItemId	*uniqueidentifier*	Straipsnio ID

6 lentelė. *DB lentelės „Period_Stat“ aprašymas*

Periodų svorių koeficientams saugoti, kai jie pasikeičia		
Laukas	**Tipas**	**Aprašymas**
Date	*datetime*	Koeficiento pasikeitimo data
Id	*uniqueidentifier*	Periodo Id (metai)
Weight	*real*	Periodo svoris skaičiuojant reitingą

7 lentelė. *DB lentelės „RatedItemRating_Stat“ aprašymas*

Straipsnių ir autorių reitingo reikšmės periodui kasdieninis saugojimas		
Laukas	**Tipas**	**Aprašymas**
Date	*datetime*	Datai kasdien saugoti
PeriodId	*uniqueidentifier*	Periodo Id (metai)
Value	*real*	Reitingo reikšmė intervale (0-1)
RatedItemId	*uniqueidentifier*	Autoriaus arba straipsnio ID

8 lentelė. *DB lentelės „RatedItemRatingElements_Stat“ aprašymas*

Straipsnių ir autorių reitingo sudėtinių elementų reikšmių periodui kasdieninis saugojimas		
Laukas	**Tipas**	**Aprašymas**
Date	*datetime*	Datai kasdien saugoti
PeriodId	*uniqueidentifier*	Periodo Id (metai)
RatedItemId	*uniqueidentifier*	Autoriaus arba straipsnio ID
RatedItemElementId	*uniqueidentifier*	Reitingo sudėtinio elemento ID
Value	*real*	Reitingo reikšmė intervale (0-1)
Number	*real*	Veiksmų kiekis kažkuriam elementui

9 lentelė. *DB lentelės „Readings_Stat“ aprašymas*

Straipsnių skaitymų kiekiams saugoti kasdien		
Laukas	**Tipas**	**Aprašymas**
Date	*datetime*	Datai kasdien saugoti
ItemId	*uniqueidentifier*	Straipsnio ID
Cnt	*int*	Skaitymų kiekis per dieną

10 lentelė. *DB lentelės „Articles_Stat“ aprašymas*

Autorių kiekviena kalba sukurtų straipsnių kiekiams saugoti kasdien		
Laukas	**Tipas**	**Aprašymas**
Date	*datetime*	Datai kasdien saugoti
Lng	*smallint*	Kalbos kodas
AuthorId	*uniqueidentifier*	Autoriaus ID
Cnt	*int*	Publikuotas straipsnių kiekis per dieną

3.3 Duomenų transformacijos

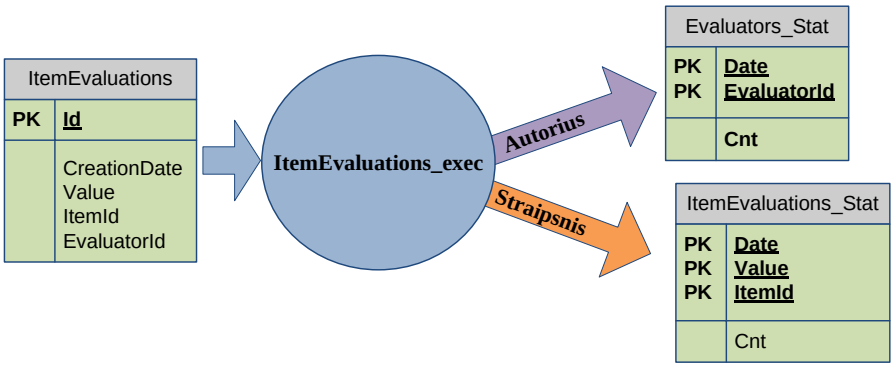
Statistikos posistemis naudojasi visos sistemos duomenų baze, bei saugo diskretizuotus duomenis į tik statistikai skirtas lenteles. Diskretizavimas padeda spręsti problemas, kur duomenys plaukia nenutrūkstamai, kaip daugelyje naujausių duomenų gavybos sistemų [20]. Kuriamame modulyje duomenys gali būti nutrūkstantys, tačiau gan intensyvūs. Visa smulki informacija nėra reikalinga statistiniam analizavimui, todėl duomenys apdorojami.

Duomenų diskretizavimas ir sumavimas vyksta serveryje saugomų procedūrų pagalba, kurios apdoroja duomenis iš bendrų portalo lentelių ir saugo juos į statistikos lenteles. Serverio valdymo įrankiuose sukuriama užduotis (job), kuri kasdien kreipiasi į portalo veiklos duomenis sumuojančias procedūras.

Portalo veiklos stebėjimui, veiksmų/įvykių analizei visiškai pakanka diskretiniais intervalais (dienomis) apibendrintos informacijos. Nagrinėti įvykius valandomis ar smulkiau visiškai nereikia, nes pati portalo veikla ir jo duomenų analizė nėra kritinio svarbumo kaip medicininių, akcijų rinkos ar pan. duomenų analizė. Tikslas yra pateikti informaciją analizei apie bendrą portalo veiklą, todėl priimtas sprendimas apjungti visos dienos informaciją ir dieną laikyti mažiausiu nedalomu laiko vienetu statistikos duomenų peržiūroje. Procedūros duomenis apdoroja dažniausiai dviem aspektais: straipsnio ir autoriaus.

3.3.1 Vertinimo duomenų apdorojimas

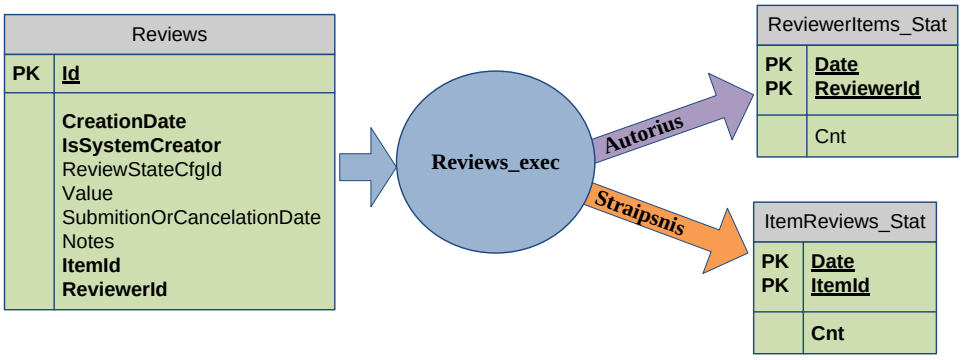
Procedūra „*ItemEvaluations_exec*“ (***15 pav.***) pagal “*ItemEvaluations*” lentelės duomenis apibendrina dienos duomenis ir informaciją autoriaus požiūriu saugo į lentelę *“Evaluators_Stat”(**2 lentelė**)*, straipsnio požiūriu – į *“ItemEvaluations_Stat”(3 **lentelė**)*.



15 pav. *Procedūros „ItemEvaluations_exec“ naudojamos/pildomos duomenų lentelės*

3.3.2 Recenzavimo duomenų apdorojimas

Procedūra „*Reviews_exec*“ (***16 pav.***) pagal “*Reviews*” lentelės duomenis apibendrina dienos duomenis ir informaciją autoriaus požiūriu apie jo recenzijų kiekius per dieną saugo į lentelę *“ReviewerItems_Stat”(**4 lentelė**)*, straipsnio požiūriu, apie straipsniui parašytas recenzijas per dieną, – į *“ItemReviews_Stat”(**5 lentelė**)*.

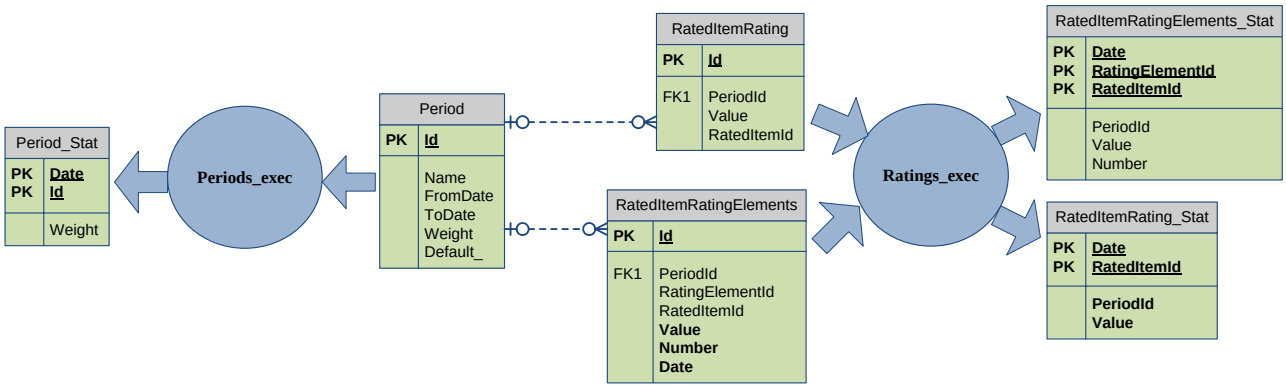


16 pav. *Procedūros „Reviews_exec“ naudojamos/pildomos duomenų lentelės*

3.3.3 Reitingavimo duomenų apdorojimas

Procedūra „*Periods_exec*“ (***17 pav.***), “*Period*” lentelėje pasikeitus einamojo periodo svoriui (tai būna labai retai), duomenis saugo į *„Period_Stat“* lentelę (***6 lentelė***).

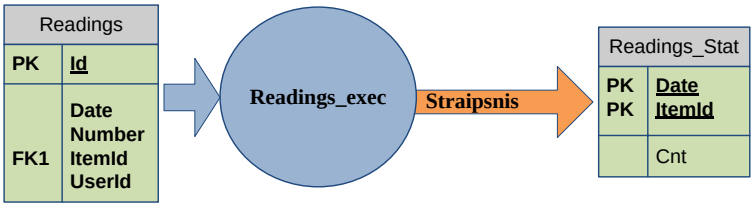
Procedūra „*Ratings_exec*“ (***17 pav.***) iš lentelės *„RatedItemRating“* einamajame periode sukauptas autorių/straipsnių reitingų reikšmes kasdien saugo į *„RatedItemRating_Stat“* lentelę (***7 lentelė***). Iš lentelės *„RatedItemRatingElements“* einamajame periode sukauptas autorių/straipsnių reitingų sudėtinių elementų reikšmes kasdien saugo į *„RatedItemRatingElements_Stat“* lentelę (***8 lentelė***).



17 pav. *Procedūrų „Periods_exec“, „Ratings_exec“ naudojamos/pildomos duomenų lentelės*

3.3.4 Straipsnių skaitymo duomenų apdorojimas

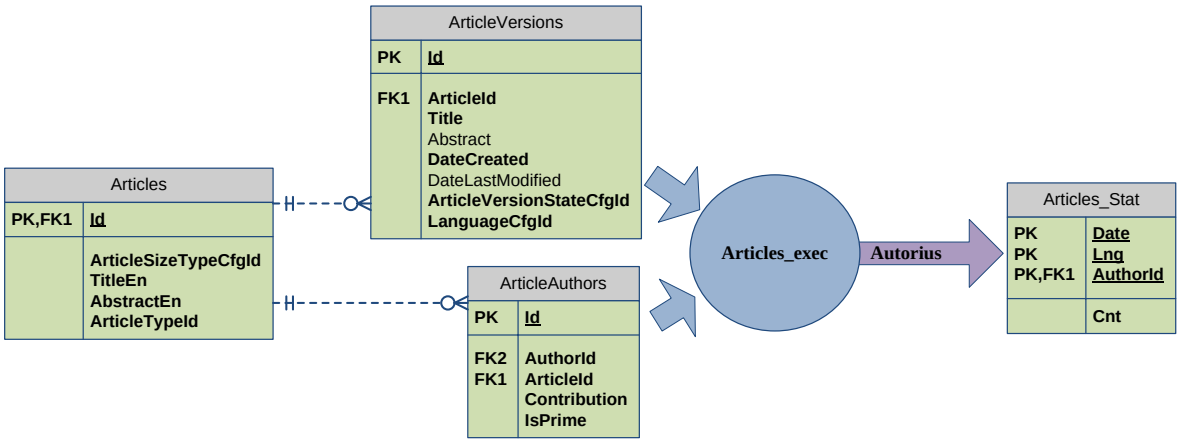
Procedūra *„Readings_exec“* (***18 pav.***) iš lentelės *„Readings“* apibendrina dienos duomenis ir informaciją straipsnių požiūriu apie jo skaitymų kiekius per dieną saugo į *„Readings_Stat“* lentelę (***9 lentelė***). Šioje vietoje apie autorių atliktų skaitymų kiekius per dieną informacija nėra aktuali, todėl nesaugoma.



18 pav. *Procedūros „Readings_exec“ naudojamos/pildomos duomenų lentelės*

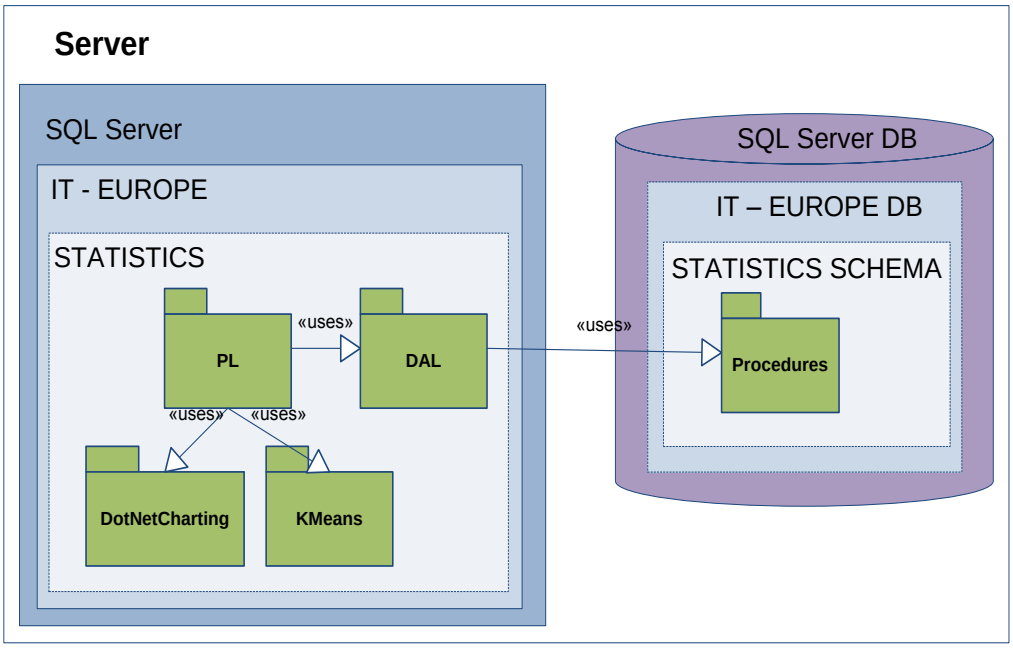
3.3.5 Straipsnių rašymo duomenų apdorojimas

Procedūra *„Articles_exec“* (***19 pav.***) iš lentelės *„Readings“* apibendrina dienos duomenis ir informaciją autorių požiūriu sumuoja kiekviena kalba straipsnių kiekius, prie kurių kūrimo jis prisidėjo, saugo į lentelę *„Articles_Stat“* (***10 lentelė***).



19 pav. *Procedūros „Articles_exec“ naudojamos/pildomos duomenų lentelės*

3.4 *Statistikos modulio architektūra*



20 pav. *Statistikos modulio architektūra*

20 pav. matomos aukščiausiame lygyje statistikos modulį sudarančios penkios dalys (komponentai):

- **„KMeans“** – Grupavimo komponentas;
- **„DotNetCharting“** – Komponentas grafiniam atvaizdavimui;
- **„PL“** – Atvaizdavimo ir valdymo dalis;
- **„DAL“** – Prieigos prie duomenų lygis;
- **„Procedures“** – Serveryje saugomos procedūros darbui su duomenimis;

Išnaudotas MS.NET Framework 2 siūlomas privalumas atskirti atvaizdavimo lygį (PL), duomenų prieigos lygį (DAL) bei serveryje saugomas procedūras.

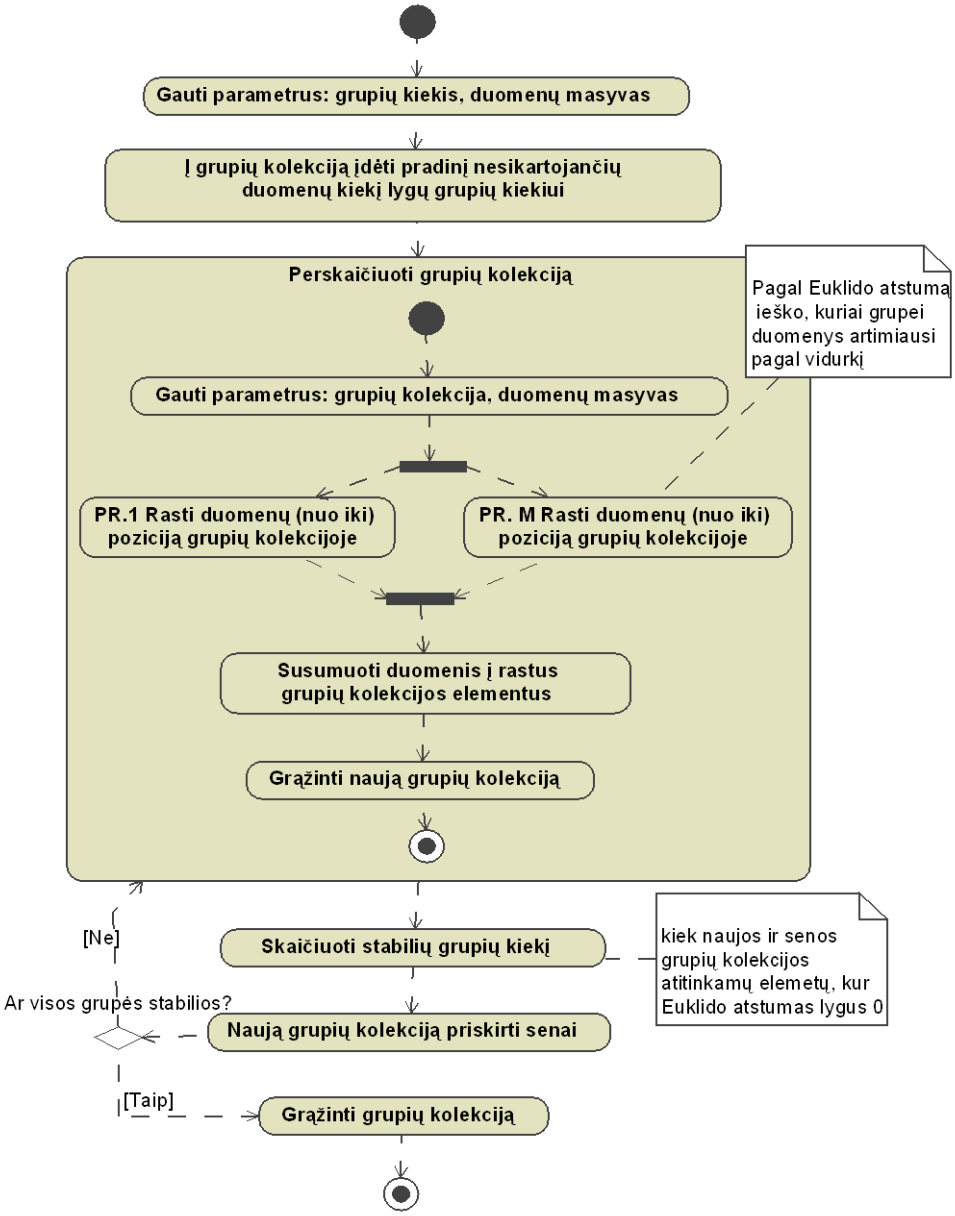
3.4.1 Komponento „KMeans“ detalizavimas

Programavimo inžinerija anksčiau buvo orientuota į originalų programinės įrangos kūrimą, bet buvo pripažinta, kad norint geriau, pigiau ir greičiau sukurti PĮ, reikia taikyti projektavimo procesą, pagrįstą pakartotiniu panaudojimu [21]. Greitesniam PĮ vystymui pasiekti pasinaudojau pakartotiniu komponentų panaudojimu. Panaudotas jau realizuotas lygiagretusis k- vidurkių algoritmas, realizuotas C# programavimo kalba [22].

Naujai bandžiau pažvelgti į anksčiau įvardintą paprastojo k- vidurkių algoritmo trūkumą dėl pradinių centrų parinkimo geresnės strategijos, kad algoritmui reikėtų mažiau iteracijų. Todėl algoritmas buvo

papildytas. Strategija – pradiniams grupių centrams iš eilės imti duomenų rinkinio nesikartojančius duomenis.

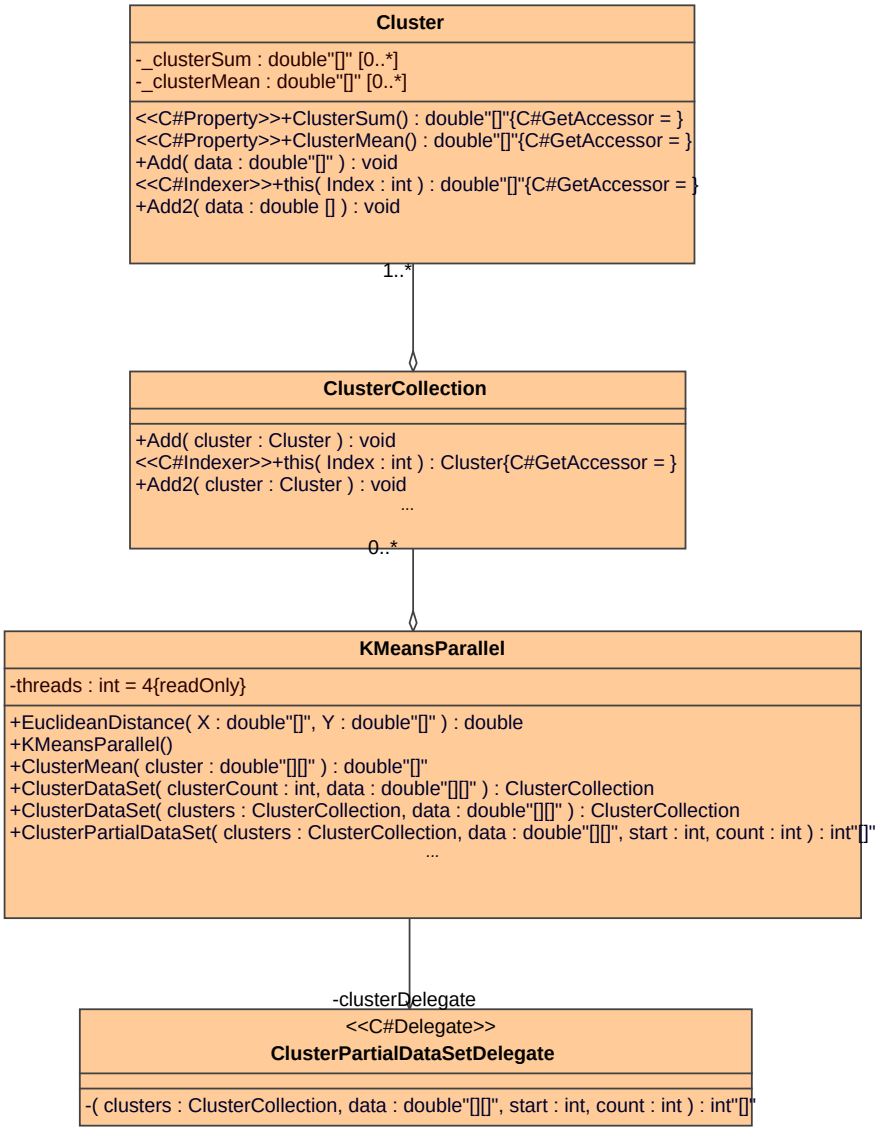
Komponentą sudaro klasės, realizuojančios kelių matų duomenų grupavimą k-vidurkių algoritmu, pagal Euklido atstumą. (***21 pav.***) Pateikiamas „*Kmeans*“ komponento duomenų rinkinio grupavimo pagrindinio metodo veikimas.



21 pav. *Duomenų rinkinio grupavimo veikos diagrama*

22 pav. parodyta grupavimo objektai, kur funkcionuoja objektas „*Cluster“*, t.y. grupė, kuri turi visų į ją patenkančių duomenų vidurkį ir sumą. Klasė „*KMeansParallel“* sukuria grupių kolekciją, jai priskiriamos pradinės reikšmės. Duomenų rinkinys išdalinamas lygiagretiems procesams. Procesai kreipiasi į delegatą „*ClusterPartialDataSetDelegate“*, kuris pagal Euklido atstumą nustato, kuriai grupei duomenys artimesni. „*KMeansParallel“* surenka duomenis iš lygiagrečių procesų ir pagal tai perskaičiuoja grupių kolekcijos

centrus.



22 pav. *Paketo „Kmeans" klasių diagrama*

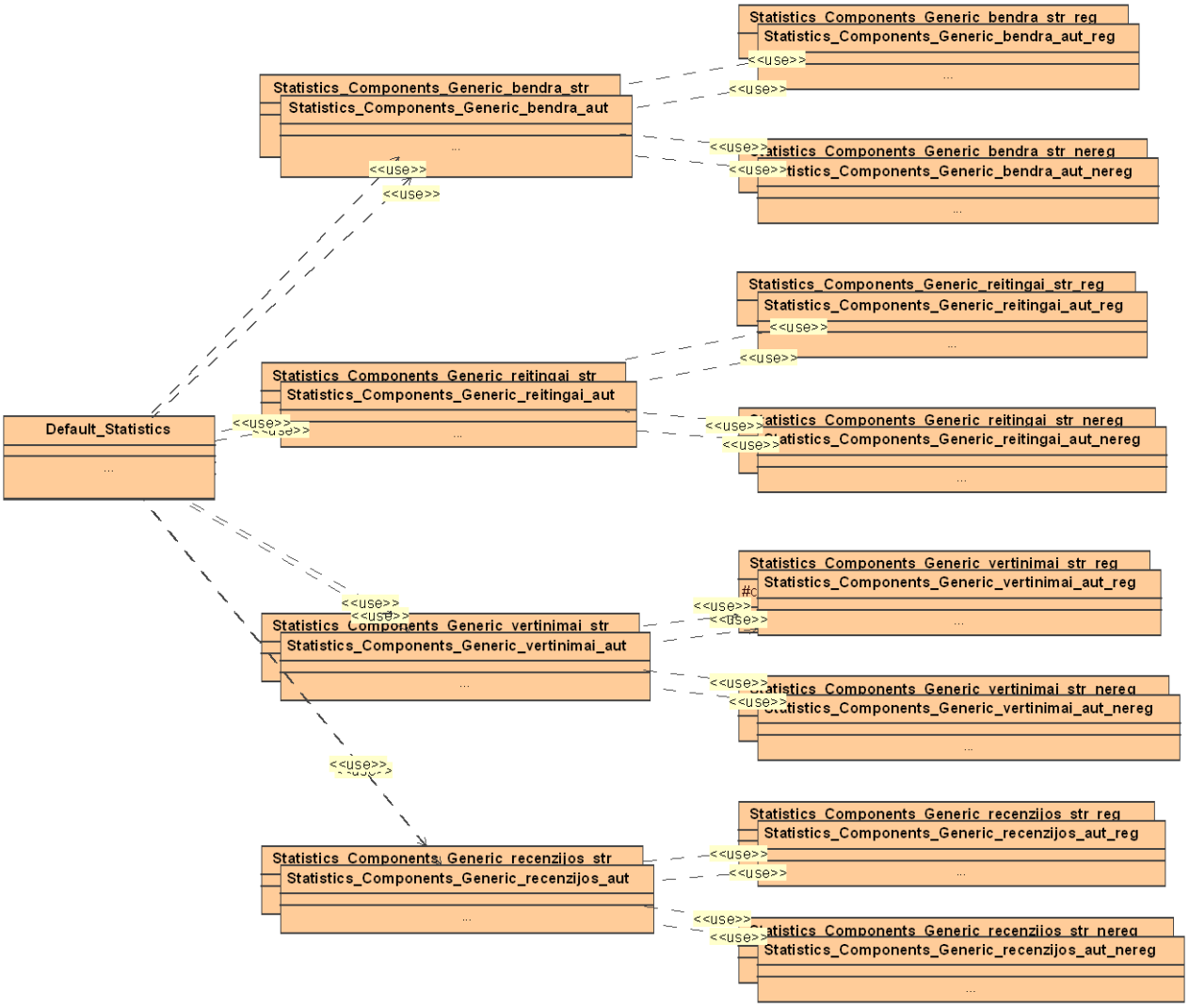
3.4.2 Komponento „DotNetCharting" detalizavimas

Panaudotas grafinio atvaizdavimo komponentas „*DotNetCharting*" (WebAvail Porductions, Inc. & Corporate Web Solutions Ltd.) [20]. Realizuotas .NET(C#) priemonėmis. Patogu naudoti ASP.NET(C#/VB) ar Windows formų programose.

Turi duomenų paėmimo galimybes tiesiogiai iš duomenų bazės (MySQL, ACCESS, SQL Server, Oracle, ODBC), Excel ar XML failų bei programinių objektų (DataSet, DataTable, DatView, ArrayList, Object Collection, HashTable). Komponentas naudojamas kaip .DLL biblioteka. Kreipiamasi į objektą, jam perduodami duomenys, gaunamas duomenų atvaizdavimas diagrama.

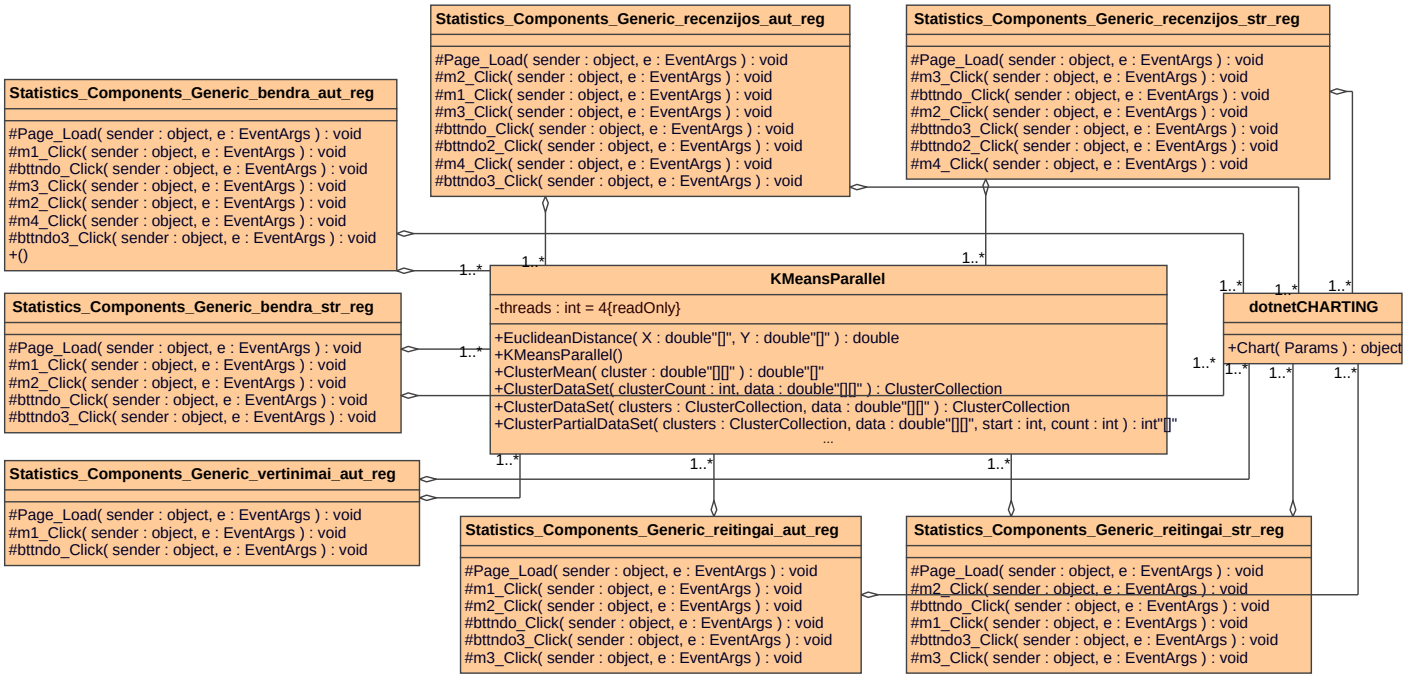
3.4.3 Komponento „PL“ detalizavimas

Pakete pateikiamos klasės, realizuojančios registruoto ir neregistruoto lankytojo sąsają, bendrą puslapių struktūrą, kitų komponentų valdymą. Paketo diagrama pateikta ***23 pav***.



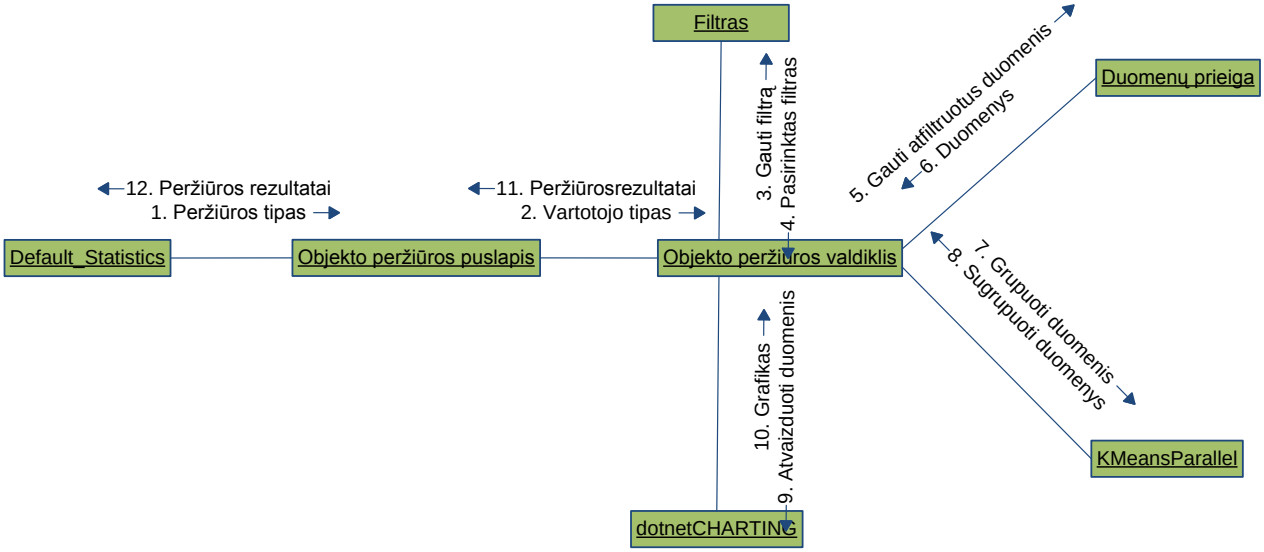
23 pav. *Paketo „PL” bendroji klasių diagrama*

Pagrindinis puslapis kreipiasi į vartotojo pasirinktą objekto dalies peržiūros puslapį (.aspx), kuris pagal vartotojo tipą kreipiasi į atitinkamą valdiklį (.ascx). Pastarieji apdoroja filtrų sukūrimą ir pasirinkimą, kreipimąsi į serverio procedūras per „DAL“ informacijai gauti. Valdikliai pagal poreikį sukuria reikiamą kiekį „*KMeansParallel*“ bei „*DotNetCharting*“ komponentų, kreipiasi į juos atitinkamai informacijai grupuoti arba grupuotai informacijai atvaizduoti. Sudėtingesnių valdiklių klasės parodytos ***24 pav***.



24 pav. *Paketo „PL" klasių diagrama grupavimui ir atvaizdavimui*

Iš aukščiau pateiktos klasių diagramos seka, bendradarbiavimo diagramos principas, tinkantis visiems panaudojimo atvejams, kur naudojamas grupavimas ir atvaizdavimas, pavaizduotas ***25 pav***. „*Default_statistics*" kviečia straipsnių arba autorių peržiūros komponentą, o šis kreipiasi į žemesnį peržiūros tipo valdiklį, kuris valdo filtro sukūrimą, atfiltruotų duomenų paėmimą, grupavimą (jei reikia) ir atvaizdavimą grafiku.

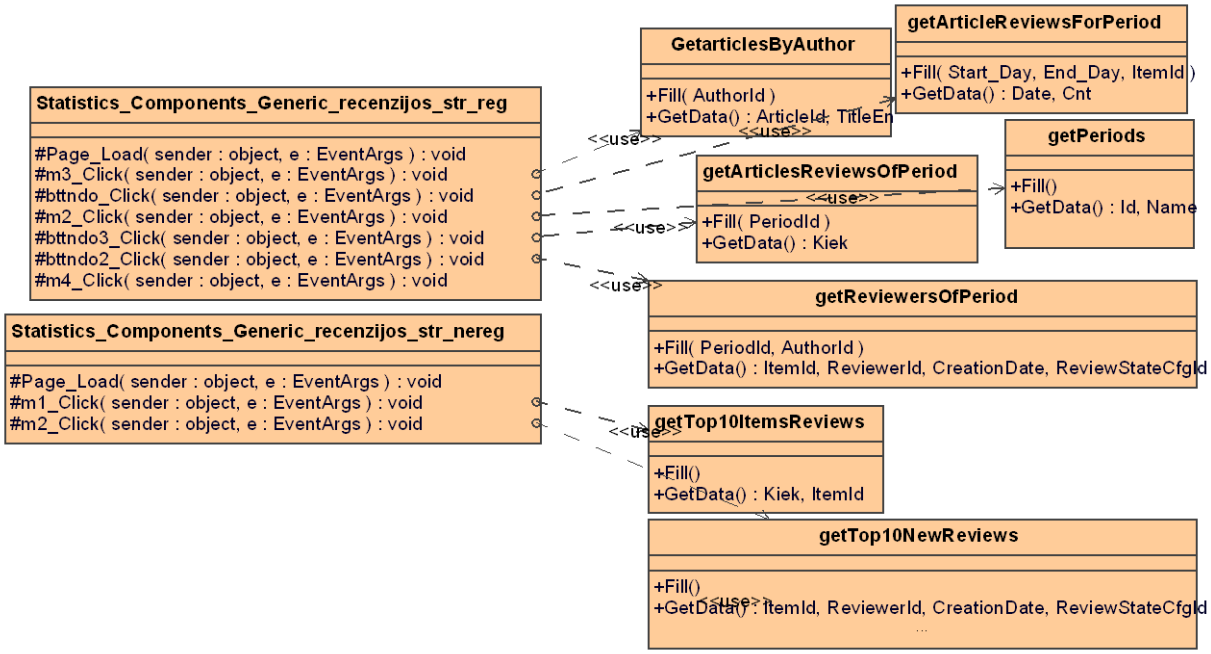


25 pav. *Apibendrinta bendradarbiavimo diagrama grupavimui ir atvaizdavimui*

3.4.4 Komponento „DAL“ detalizavimas

„DAL“ – Duomenų prieigos lygis. Čia sukauptas visas specifinis kodas, skirtas prieiti duomenims iš duombazės bei prisijungimo prie duombazės nustatymai. „*DataSet“* objektas apima „*DataTables“*, „*TableAdapters“*. „*TableAdapter“* – klasė su metodais darbui tarp „*DataSet“* objekte esančių „*DataTable“* ir duomenų bazės. „*TableAdapter“* metodas „*Fill()“* užpildo „*DataTable“* lentelę, o pakeistus ar naujus duomenis siunčia į duombazę naudojant metodus („*Insert()“, „Update()“, „Select()“, „Delete()“*), skirtų darbui su reikiamos duombazės lentelės duomenimis. Minėti metodai perduoda reikiamus parametrus duombazėje saugomoms atitinkamos paskirties procedūroms [23].

Atvaizdavimo lygis „PL“ turi kreiptis į „DAL“ metodus darbui su norimos lentelės duomenimis. (***26 pav.***). Pateikiamas registruotų ir neregistruotų vartotojų straipsnių recenzijų informacijos peržiūros klasių („*Statistics_Components_Generic_recenzijos_str_reg*“, „*Statistics_Components_Generic_recenzijos_str_nereg*“) kreipimasis į „*TableAdapter“* klases duomenims paimti.



26 pav. *„DAL“ klasių diagramos pavyzdinė dalis straipsnių recenzijoms*

3.4.5 Komponento „Procedures” detalizavimas

„Procedures“ – tai MSSQL serverio duomenų bazėje saugomos SQL procedūros. Saugomų procedūrų nauda:

- Sukompiliavimas iš anksto atsiperka, kai procedūros naudojamos pakartotinai;
- Sumažintas apkrovimas serveryje;
- Gali būti naudojamos kelių programų ar vartotojų;
- Padidinta saugumo kontrolė leidžia suteikti teises vartotojams vykdyti procedūras nepriklausomai nuo lentelės teisių; [24]

„DAL“ turi prieigą prie serverio procedūrų, kurios naudojamos komponente „PL“ darbui su duomenimis. Procedūrų paaiškinimus galima rasti ***priede***.

4 STATISTIKOS MODULIO REALIZACIJA

4.1 Statistikos modulio veikimas

Statistikos modulio sąsaja parengta pagal funkcionalumo lentelės struktūrą (žiūrėti **2.1.4** skyriaus ***1 lentelę***). Sistemos naudojimas nėra sudėtingas. Keletas patarimų nesusipratimams išvengti:

- Pasirinkus laiko intervalą gali jame duomenų nebūti, tada matomas pranešimas „No Data“. Sprendimas būtų bandyti rinktis platesnį ar visai kitą laiko intervalą;
- Laiko intervalą pasirinkus per didelį, informacija gali būti sunkiai suprantama. Reikia bandyt mažinti laiko intervalą;
- Pasirinkus grupių kiekį, gali nebūti tiek skirtingų duomenų (pvz.: visi straipsniai skaityti po 1 kartą), tada gaunama tik tiek grupių, kiek randama skirtingų duomenų;

4.1.1 Lankytojams prieinamos funkcijos

- **Straipsnių statistikos peržiūros galimybės**
  - **Bendra statistika**

Skirta apibendrintai informacijai apie straipsnių skaitymus ir parašymus peržiūrėti (***27 pav.***).

Galima pasirinkti peržiūrą pateikiančią einamajame periode dešimties skaitomiausių straipsnių informaciją (straipsnis, sukūrimo data, skaitymų kiekis pamatyti). Taip pat galima peržiūrėti dešimt naujausių straipsnių (straipsnis, sukūrimo data). Pateikiama paskutinio periodo kiekvieno mėnesio straipsnių skaitymų informacija (mėnuo, skaitymų kiekis) bei einamajame periode kas mėnesį parašytų straipsnių informacija (mėnuo, straipsnių kiekis). Taip pat galima matyti bendras straipsnių kiekį sistemoje.

AUTORIAI | STRAIPSNIAI

Bendra | Recenzijos | Vertinimai | Reitingai

HOME > Statistics

Bendra straipsnių informacija

< Atgal

▪ Top10 skaitomiausių straipsnių einamam periode

Nr.	Straipsnis	Sukurtas	Skaitymų kiekis
1	RST akcijos bus įtrauktos į Oficialųjį prekybos sąrašą	2007-01-19	60
2	"Dow Jones" toliau auga, euro stiprėjimas baigiasi	2007-01-10	40
3	Ką suteikia Įmonės Tinklapis?	2007-01-22	35
4	Analitikai rekomenduoja parduoti „Baltika" akcijas	2007-01-08	31
5	Antradienį rinka išskaičiavo dividendus iš TEO akcijų kainos	2007-01-18	30
6	Ketvirtadienį biržoje dominavo pasyvumas	2007-01-07	29
7	Ekonomistai spėlioja, kada sustos JAV darbo vietų gamybos mašina	2007-01-08	25
8	Microsoft tikslina verslo analizės produktų žemėlapį	2007-03-17	18
9	Microsoft Office 2007 suderinamumas su Microsoft Dynamics NAV 4.0 SP3	2007-03-11	17
10	Išleista MagicDraw UML 12.1 versija	2007-02-17	15

▪ Top10 naujausių straipsnių

Nr.	Straipsnis	Sukūrimo data
1	Kas naujo SQL serveryje 2005? Penki būdai, kaip gauti daugiausia naudos iš naujos duomenų bazės versijos	2007-04-22
2	Microsoft Windows Server - geriausiai parduodama serverio operacinė sistema	2007-04-15
3	Verslo analizės (BI) įrankiai ne vien didelėms įmonėms	2007-04-11
4	Microsoft planuoja daugiau "snap" sąsajų tarp "Microsoft Office" ir VVS sistemų	2007-04-08
5	Kaip toli Microsoft gali eiti?	2007-03-20
6	Microsoft tikslina verslo analizės produktų žemėlapį	2007-03-17
7	UAB "Sonex sistemos" tapo Auksine Sertifikuota Microsoft Partnere	2007-03-15
8	„Microsoft Dynamics NAV" jau naudojasi milijonas vartotojų	2007-03-15
9	Microsoft Office 2007 suderinamumas su Microsoft Dynamics NAV 4.0 SP3	2007-03-11
10	Sujunkite „Microsoft Dynamics NAV" su „Microsoft Dynamics CRM"	2007-03-11

▪ Einamam periode kas mėnesį parašytų straipsnių kiekiai

Mėnuo	Skaitymų kiekis
1	49
2	55
3	302
4	1284

▪ Einamam periode kas mėnesį parašytų straipsnių kiekiai

Mėnuo	Publikuotų straipsnių kiekis
1	17
2	8
3	12
4	4

27 pav. *Bendra straipsnių informacija lankytojui*

- **Recenzijų statistika**

Skirta bendrai informacijai apie straipsnių recenzijas peržiūrėti (***28 pav.***). Galima pasirinkti dešimties daugiausiai einamajam periode recenzuotų straipsnių sąrašą (straipsnis, recenzijų kiekis) arba dešimties naujausiai recenzuotų straipsnių sąrašą (straipsnis, recenzentas, recenzijos data).

AUTORIAI | STRAIPSNIAI

Bendra | Recenzijos | Vertinimai | Reitingai

HOME > Statistics

Straipsnių recenzijų informacija < Atgal

• Top10 labiausiai recenzuotų straipsnių einamajam periode

Nr.	Straipsnis	Recenzijų kiekis
1	Programinės įrangos defektų valdymas	2
2	Lietuvoje trūksta geografinių informacinių sistemų specialistų	2
3	Microsoft Office 2007 suderinamumas su Microsoft Dynamics NAV 4.0 SP3	1
4	Per „Skype" plintantis virusas ištuštino ne vieną sąskaitą	1
5	Pristatyta nauja „Windows Server" bandomoji versija	1
6	Lietuvoje trūksta geografinių informacinių sistemų specialistų	1
7	Išleista MagicDraw UML 12.1 versija	1
8	Sonex Sistemos dalyvavo Microsoft CEE Partnerių susitikime Budapešte	1
9	„Bitė" pristatė virtualią ugniasienę verslui	1
10	Sonex Sistemos dalyvavo Microsoft konferencijoje klientams "IT sprendimai Jūsų verslui"	1

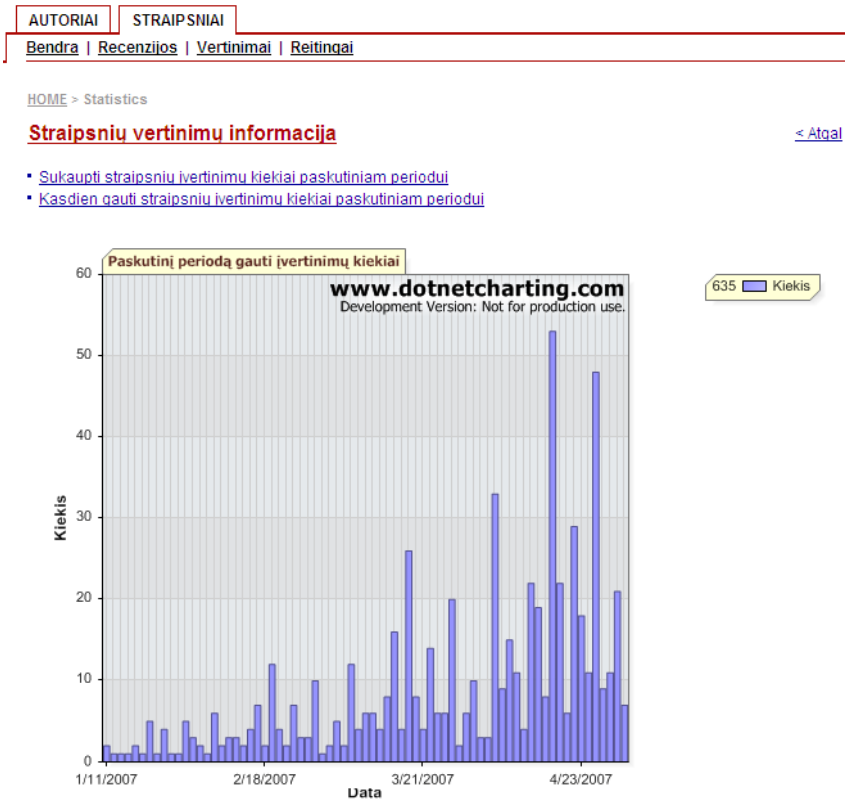
• Top10 naujausiai recenzuotų straipsnių

Nr.	Straipsnis	Recenzentas	Recenzijos data
1	ITG teikiamų paslaugų kokybei suteiktas ISO 9001:2000 sertifikatas	Markas Markevičius	2007-01-04
2	ITG aktyvina veiklą Estijoje	Kęstas Kęstaitis	2007-01-06
3	Lietuvoje trūksta geografinių informacinių sistemų specialistų	Modestas Modestaitis	2007-01-08
4	Per „Skype" plintantis virusas ištuštino ne vieną sąskaitą	Inga Ingaitė	2007-01-11
5	"Dow Jones" toliau auga, euro stiprėjimas baigiasi	Rūta Rūtaitė	2007-01-21
6	Pristatyta nauja „Windows Server" bandomoji versija	Kęstas Kęstaitis	2007-01-22
7	Antradienį rinka išskaičiavo dividendus iš TEO akcijų kainos	Jonas Jonaitis	2007-01-27
8	RST akcijos bus įtrauktos į Oficialųjį prekybos sąrašą	Linas Linaitis	2007-01-27
9	Akcijų pirkimas-pardavimas ar įmonės pirkimas-pardavimas – ką rinktis?	Jonas Jonaitis	2007-01-29
10	Ką suteikia Įmonės Tinklapis?	Radvila Radvilaitė	2007-01-29

28 pav. *Bendra straipsnių recenzijų informacija lankytojui*

- **Vertinimų statistika**

Skirta bendrai informacijai apie straipsnių vertinimus peržiūrėti (**29 *pav.***). Galima peržiūrėti per paskutinį periodą kiekvieną dieną gautus straipsnių vertinimų kiekius arba iki kiekvienos dienos sukauptus straipsnių vertinimų kiekius.



29 pav. *Bendra straipsnių vertinimo informacija lankytojui*

- **Reitingų statistika**

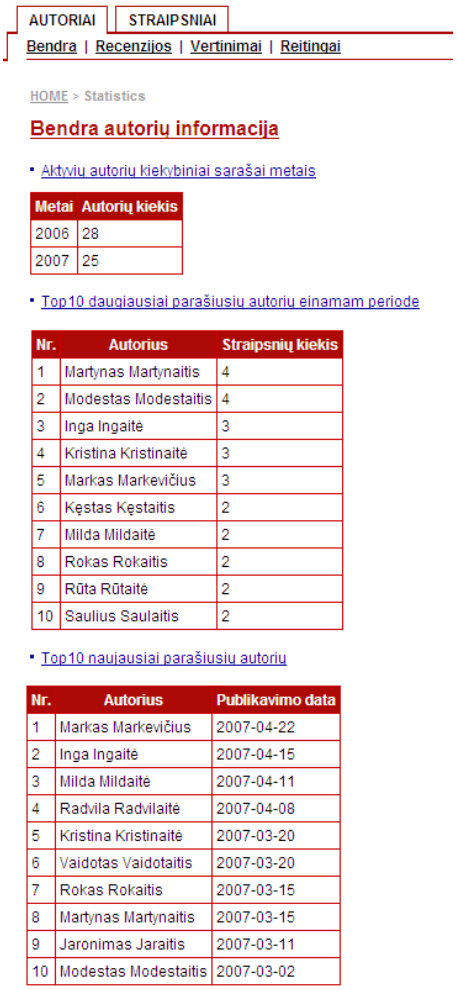
Skirta bendrai informacijai apie straipsnių reitingus peržiūrėti. Galima matyti einamajame periode dešimt aukščiausių reitingų pasiekusių straipsnių sąrašą (straipsnis, reitingo reikšmė) ir kt.

❖Autorių statistikos peržiūros galimybės

- **Bendra statistika**

Skirta bendrai informacijai apie autorius peržiūrėti (***30 pav.***).

Pasirinkus autorių bendros statistikos peržiūrą, galima matyti aktyvių autorių kiekius kiekvienais metais. Galima pasirinkti peržiūrėti dešimt daugiausiai parašiusių autorių einamajam periode (autorius, jo parašytų straipsnių kiekis). Taip pat galima peržiūrėti dešimt naujausiai rašiusių autorių (autorius, publikavimo data).



30 pav. *Bendros autorių informacijos peržiūra lankytojams*

- **Recenzijų statistika**

Skirta apibendrintai informacijai apie autorių parašytas recenzijas peržiūrėti (***31 pav.***).

Galima matyti dešimt daugiausiai recenzuojančių autorių sąrašą (autorius, recenzijų kiekis) ir kt.

31 pav. *Bendros autorių recenzijų informacijos peržiūra lankytojams*

- **Vertinimų statistika**

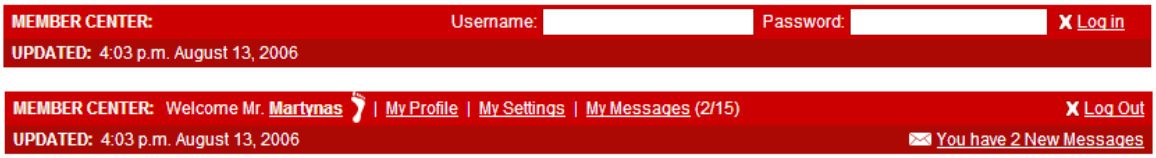
Skirta apibendrintai autorių, vertinančių straipsnius, statistikai peržiūrėti. Galima peržiūrėti dešimt daugiausiai vertinimų atlikusių autorių sąrašą ir kt.

- **Reitingų statistika**

Skirta apibendrintai autorių reitingų informacijai peržiūrėti. Galima matyti geriausius reitingus pasiekusių autorių sąrašą ir pan.

4.1.2 Registruotiems vartotojams prieinamos funkcijos

Sistemos funkcijas, prieinamas paprastiems lankytojams, gali naudoti ir sistemos autoriai (registruoti lankytojai). Prisijungęs autorius (***32 pav.***) mato visas paprasto lankytojo peržiūros galimybes bei turi papildomas toliau aprašomas galimybes.



32 pav. *Registruotų vartotojų prisijungimas*

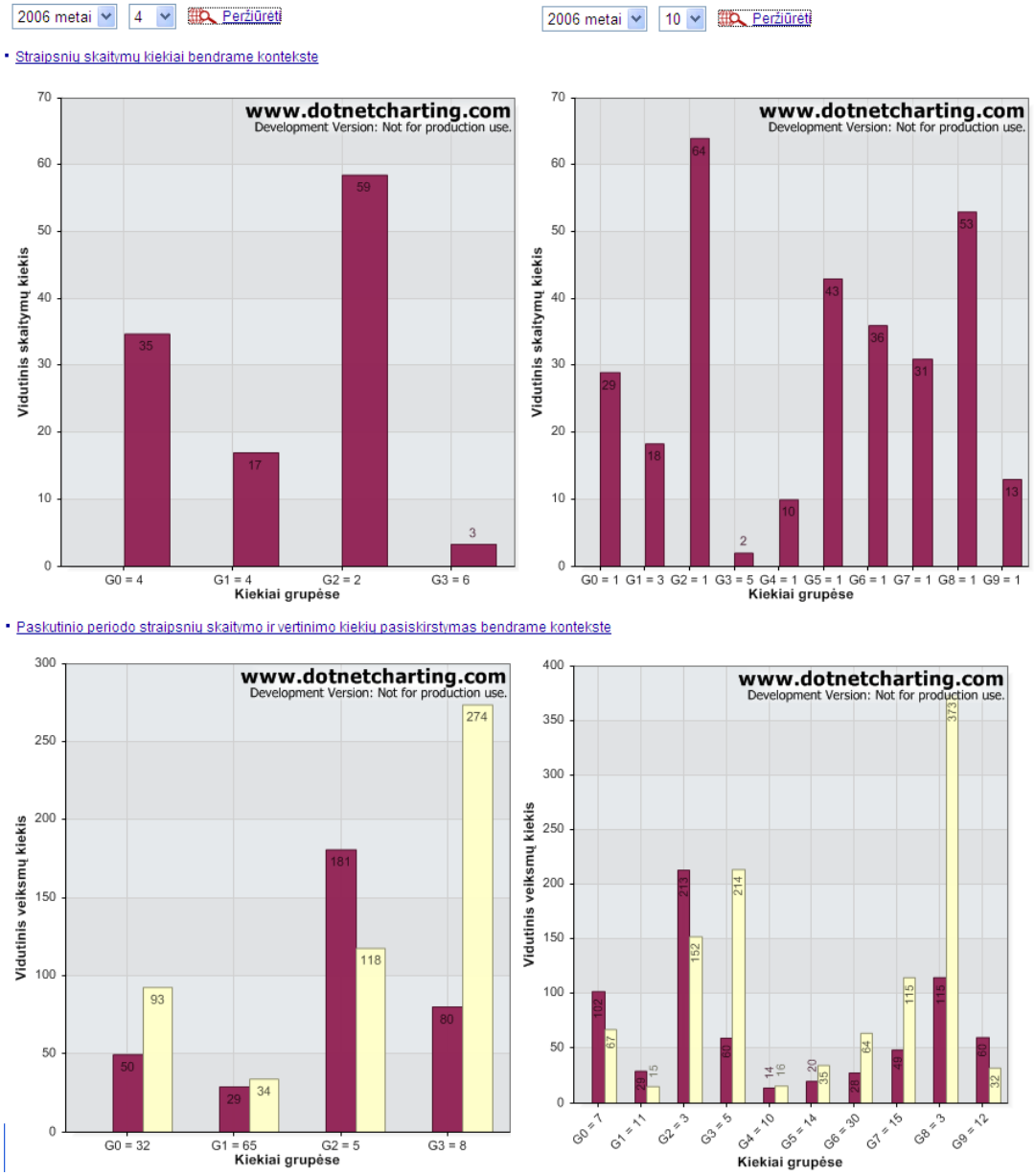
❖**Straipsnių statistikos peržiūros galimybės**

- **Bendra statistika**

Skirta detaliai straipsnių skaitymų ir rašymų statistikai peržiūrėti.

Galima pasirinkti periodą ir matyti straipsnių skaitymų kiekius bendrame kontekste arba skaitymų ir vertinimų kiekių priklausomybės pasiskirstymą (***33 pav.***).

Taip pat pasirinkus laiko intervalą galima matyti prisijungusio autoriaus parašytų straipsnių kiekius laike.



33 pav. *Detalios straipsnių informacijos peržiūra registruotiems vartotojams*

- **Recenzijų statistika**

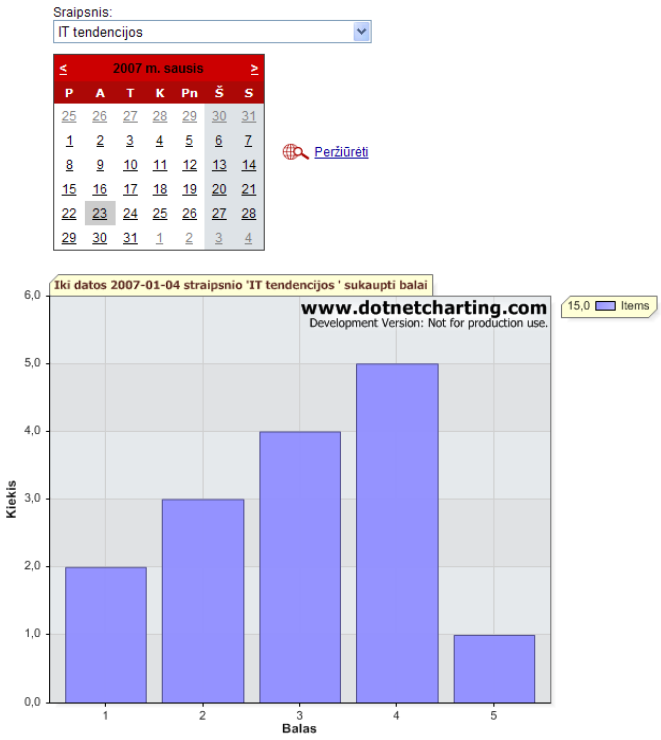
Skirta detaliai prisijungusio autoriaus parašytų recenzijų statistikai peržiūrėti.

Galima peržiūrėti autoriaus straipsnių recenzentų kiekius kasdien laiko intervale, taip pat autorių atliktų recenzijų kiekius bendrame kontekste ar nurodyto periodo autoriaus straipsnių recenzijų sąrašus (straipsnis, recenzentas, data, būsena).

- **Vertinimų statistika**

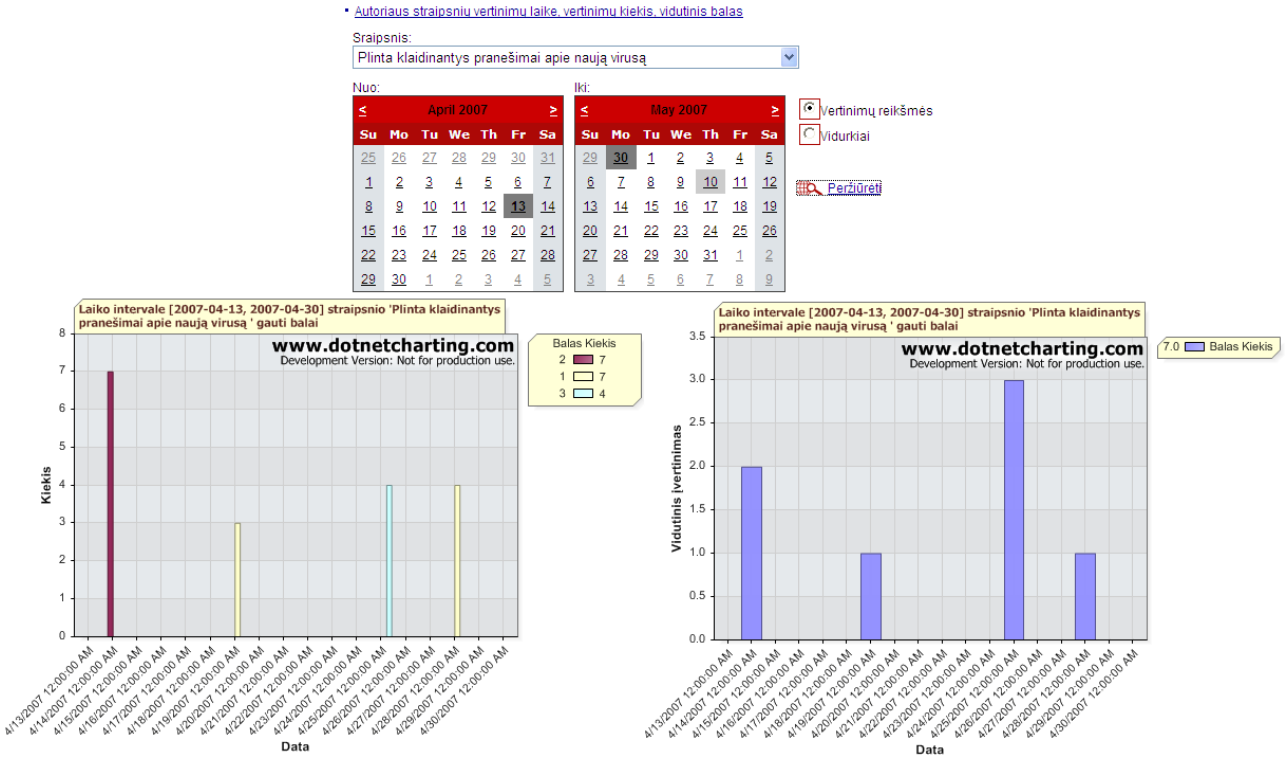
Skirta detaliai prisijungusio autoriaus straipsnių vertinimų statistikai peržiūrėti.

Pasirinkus norimą straipsnį ir datą, matomi straipsnio sukaupti balai iki pasirinktos datos (***34 pav.***).



34 pav. *Kiekvieno balo vertinimų kiekių grafikas*

Galima peržiūrėti kasdien laiko intervale norimo autoriaus straipsnio kiekvieno balo vertinimų kiekius arba vertinimų vidurkius laiko intervale kasdien (***35 pav.***).

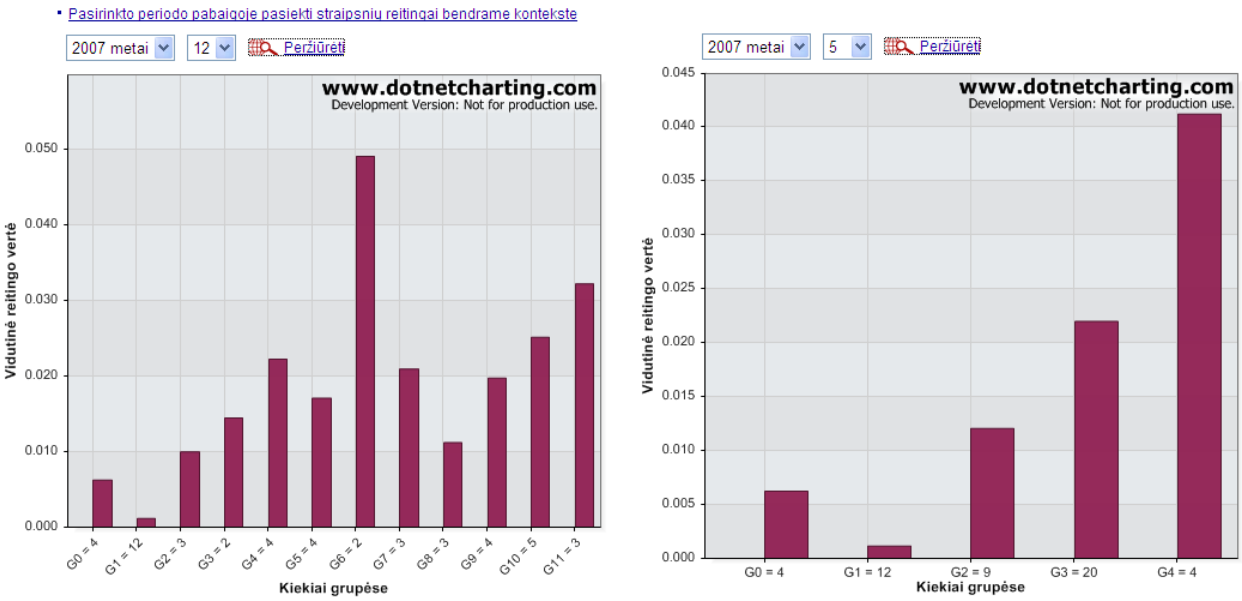


35 pav. *Autoriaus straipsnių vertinimų kiekiai laiko intervale*

- **Reitingų statistika**

Skirta detaliai autoriaus straipsnių reitingų statistikai peržiūrėti (***36 pav.***).

Galima matyti pasirinkto straipsnio reitingo vertės kitimą laiko intervale. Taip pat galima pasirinkto periodo pabaigoje straipsnių pasiektų reitingų bendrame kontekste peržiūra. Tam pasirenkamas periodas ir grupių kiekis.

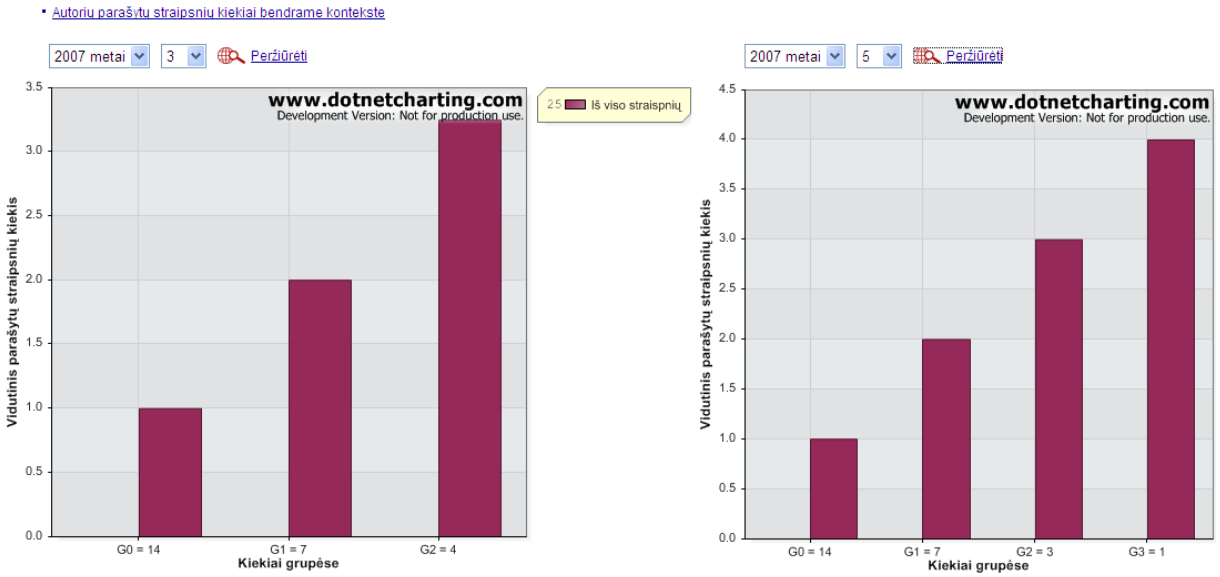


36 pav. *Straipsnių pasiekti reitingai bendrame kontekste*

❖Autorių statistikos peržiūros galimybės

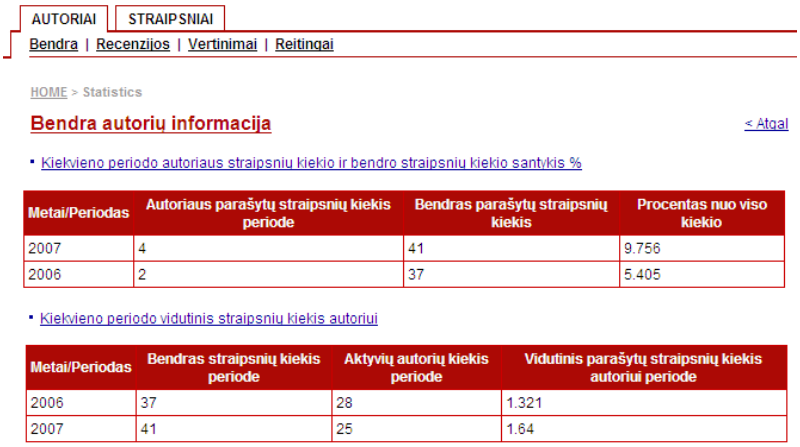
➢ Bendra statistika

Skirta detaliai autorių rašymo statistikai peržiūrėti. Galima pasirinkti periodą ir grupių kiekį, o tada matyti Bendrame kontekste autorių parašytų straipsnių kiekius per nurodytą periodą (***37 pav.***).



37 pav. *Sugrupuoti straipsnių rašymo kiekiai*

Taip pat galima matyti kiekvieno periodo autoriaus straipsnių kiekio ir bendro straipsnių kiekio santykį procentais bei vidutinius straipsnių kiekius autoriui (***38 pav.***).


AUTORIAI | STRAIPSNIAI

Bendra | Recenzijos | Vertinimai | Reitingai

HOME > Statistics

Bendra autorių informacija < Atgal

- Kiekvieno periodo autoriaus straipsnių kiekio ir bendro straipsnių kiekio santykis %

Metai/Periodas	Autoriaus parašytų straipsnių kiekis periode	Bendras parašytų straipsnių kiekis	Procentas nuo viso kiekio
2007	4	41	9.756
2006	2	37	5.405

- Kiekvieno periodo vidutinis straipsnių kiekis autoriui

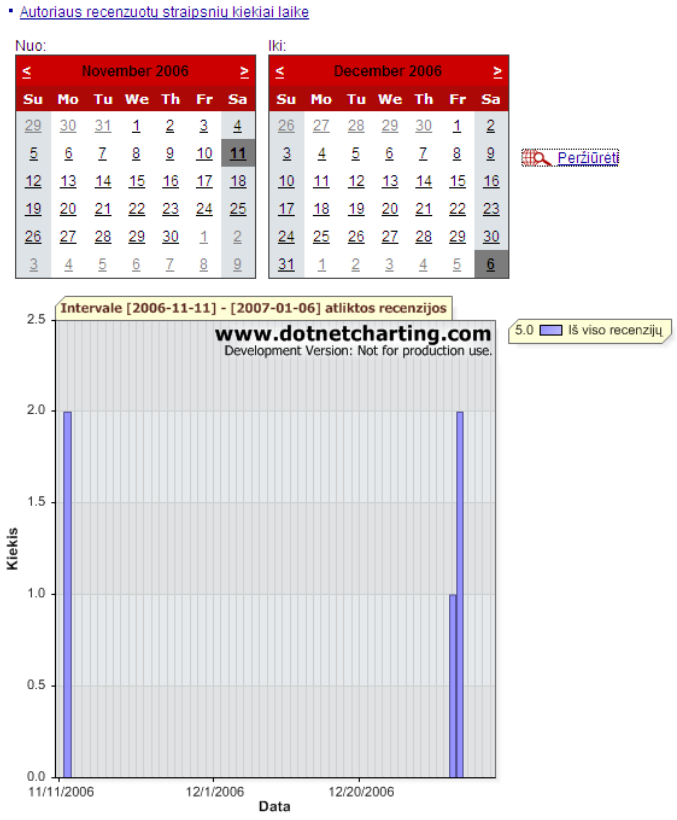
Metai/Periodas	Bendras straipsnių kiekis periode	Aktyvių autorių kiekis periode	Vidutinis parašytų straipsnių kiekis autoriui periode
2006	37	28	1.321
2007	41	25	1.64

38 pav. *Autoriaus straipsnių rašymo vidutinės ir santykinės reikšmės*

- **Recenzijų statistika**

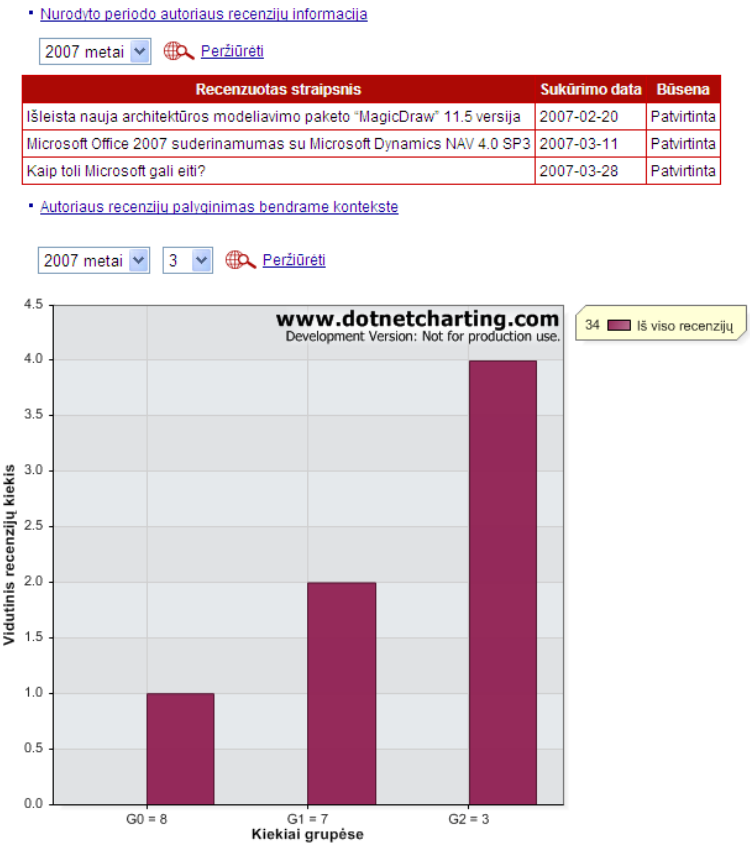
Skirta detaliai autoriaus atliktų recenzijų statistikai peržiūrėti.

Pasirinkus laiko intervalą, galima matyti kasdien autoriaus recenzuotų straipsnių kiekius intervale **(39 pav.)**.



39 pav. *Kiekiai laike*

Taip pat galima peržiūrėti nurodyto periodo autorių recenzijų kiekius sugrupuotus į norimą kiekį grupių bei pasirinkto periodo autoriaus recenzijų informaciją ***(40 pav.)***.



40 pav. *Autoriaus recenzijų informacija*

- **Vertinimų statistika**

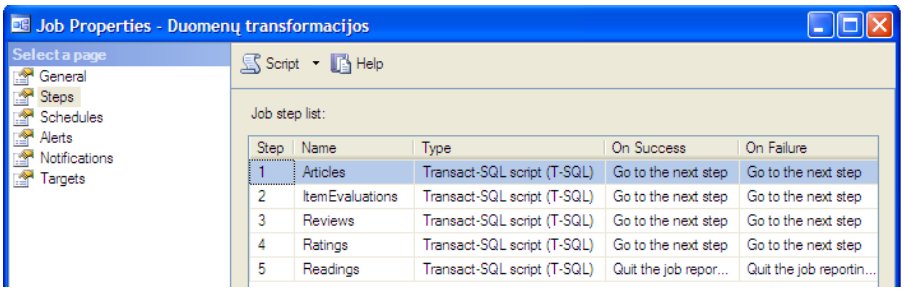
Skirta detaliai autoriaus atliktų vertinimų statistikai peržiūrėti. Gali būti peržiūrimi autorių atliktų vertinimų kiekiai sugrupuoti bendrame kontekste ir kt.

- **Reitingų statistika**

Skirta detaliai autoriaus reitingų statistikos peržiūrai. Laiko intervale galima matyti autoriaus reitingo vertės kitimą bei pasirinkto periodo pabaigoje autorių pasiektus reitingus sugrupuotus į norimą kiekį grupių.

4.1.3 Duomenų transformacijų veikimas

Duomenų transformacijoms vykdyti sukurtos procedūros: *„Articles_sp_exec“*, *„ItemEvaluations_exec“*, *„ItemRatings_exec“*, *„Periods_exec“*, *„Ratings_exec“*, *„Readings_exec“*, *„Reviews_exec“*. Joms vykdyti DB serveryje yra sukurta užduotis, sudaryta iš kelių žingsnių (***41 pav.***).

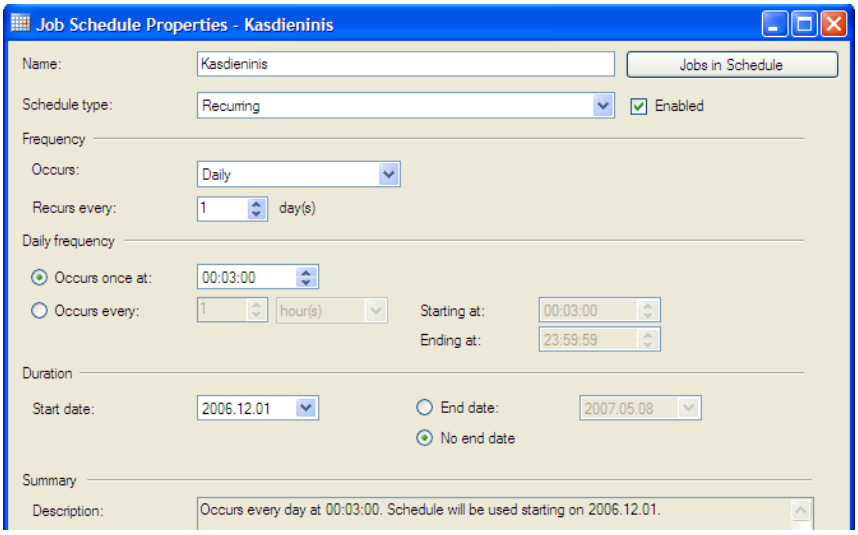


41 pav. *Duomenų transformacijos užduotis DB serveryje*

Kiekvienam užduoties žingsniui nurodoma procedūra (***42 pav.***).

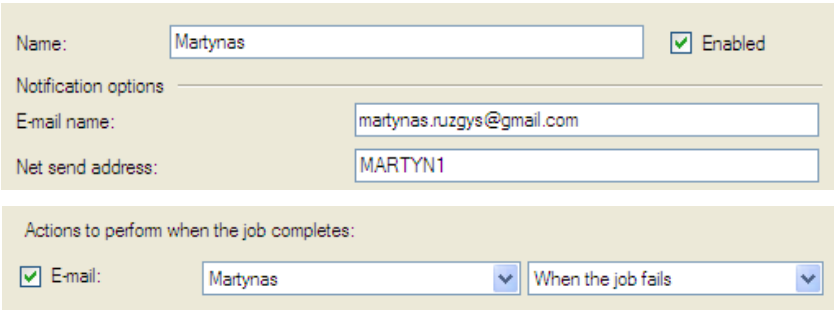
42 pav. *Užduoties žingsnio turinys*

Transformacijos vykdomos kasdien po vidurnakčio (***43 pav.***). Diskretizuojami ką tik pasibaigusios dienos sukaupti duomenys.



43 pav. *Transformacijų vykdymo grafikas*

Sukuriamas operatorius, kuriam el. paštu pranešamos klaidos (***44 pav.***).



44 pav. *Serverio užduoties pranešimų nustatymai*

4.2 Įgyvendinimo priemonių parinkimas

Statistikos modulio realizacijai naudojamą duomenų bazės valdymo sistemą (DBVS) diktuoja visas portalas. Jam kurti pasirinkta MSSQL serverio DBVS. Taip pat realizacijai parinkta Microsoft Visual Studio .Net 2005 programavimo aplinka, nes geriausiai tinka prie MSSQL serverio DBVS ir interpretuoja bet kuria iš pagrindinių programavimo kalbų parašytą tekstą, o taip pat palaiko reikalingus standartus ir puikiai tinka internetinėms aplikacijoms kurti. Pasirinkimą taip pat lemia sprendimas visą portalą realizuoti ASP.NET(C#) programavimo kalba. Bus naudojama patogi prieigos prie duomenų bazių technologija ADO.NET bei MS.NET Framework 2 siūlomas duomenų prieigos lygio atskyrimas.

4.3 Statistikos modulio realizacijos apibendrinimas

Sukurto modulio testavimui naudota stambinančioji metodika. Pirmiausiai buvo testuojamos atskiros saugomos serverio procedūros, tada programos kreipimasis į jas. Buvo rankiniu būdu įvesti straipsniai ir autoriai, o jų veiksmai generuojami Visual Studio pagalba. Ištestuotas korektiškas grupavimas ir atvaizdavimas. Pastebėtos klaidos iš karto buvo taisomos ir sistema turi veikti teisingai. Duomenų fragmentai pateikti ***priede Nr.2***.

Statistinės programinės įrangos kokybę įvertinta naudojant tokius parametrus (***11 lentelė***).

11 lentelė. *Statistikos PĮ vertinimo kriterijai*

Nr.	Parametras	Aprašymas
1.	Tikslumas	Pateikiamų skaičiavimų ir išrinktų duomenų tikslumas
2.	Sąsaja	Patogumas peržiūrėti, suprantamumas kai gausu duomenų
3.	Panaudojamumas	Lengvumas išmokti dirbti su programine įranga
4.	Patvarumas	Tolerantiškumas vartotojo klaidoms
5.	Funkcionalumas	Peržiūros tipų gausa, procentinių, palyginamųjų, vidutinių reikšmių gausa

- Produktas atitinka užsibrėžtus kokybės reikalavimus
- Produkte buvo realizuota dauguma specifikacijoje apibrėžtų funkcijų.

5 STATISTIKOS MODULIO EKSPERIMENTINIS TYRIMAS IR VERTINIMAS

5.1 Duomenų gavybos ir atvaizdavimo savybių palyginimas

Sukurtas statistikos modulis (SM) su pažangiom Statistika ir StarProbe sistemomis savo funkcionalumu nelabai sulyginamas (***12 lentelė***), nes tai yra galingos sistemos, naudojamos medicininiams ar akcijų rinkos duomenims tirti. Bet su šiomis sistemomis galima atrasti bendrų taškų DG ir grupavimo srityje, nes naudojamos artimos algoritmų realizacijos. Mineset įrankis daug kuo panašus į SM, nes jo pagrindas taip pat

k-vidurkių algoritmas. Pastarajam leidžiami atstumo mato skaičiavimo pasirinkimai tarp Euklido arba Manheteno atstumų, kas sukurtoje sistemoje lengvai gali būti padaryta, pakeičiant vieną metodą, bet realizacijoje pasirinktas vienas matas, nes matų pasirinkimai nereikalingi. SM, kaip ir Mineset leidžia pasirinkti norimą grupių kiekį, o tai ir buvo vienas iš tikslų kuriamoj sistemoj. Atvaizdavimo priemonės nagrinėtose sistemose taip pat pažangios, nors portalo statistikos pateikimui pakanka paprastu grafinių priemonių, nes tai yra internetinis sprendimas ir galingos priemonės apsunkintų veikimą.

12 lentelė. *Duomenų gavybos galimybių suvestinė*

			STATISTICA	**StarProbe**	**Mineset**	**SM**
Duomenų gavyba	**Grupavimas**	**Metodai**	k-vidurkių, k-artimiausių kaimynų, EM, medžių grupavimas, tikimybinis grupavimas	Neuroninis grupavimas, save organizuojantys žemėlapiai (SOM), segmentacija	k-vidurkių ir iteratyvusis k-vidurkių grupavimo metodas	Lygiagretusis k-vidurkių grupavimo metodas
		Matai			Euklido, Manheteno atstumas	Euklido (Manheteno) atstumas
		Pasirinkimai	v-fold maišymas ir kryžminis tikrinimas, dinaminis grupių kiekio parinkimas		Grupių kiekio, maksimalus iteracijų skaičiaus, atributų svorio pasirinkimas, atsitiktinis pradinių grupių reikšmių suteikimas	Grupių kiekio pasirinkimas (lygiagrečių procesų kiekis), pradinių grupių reikšmės – nesikartojantys duomenys iš eilės
	Prognozavimas		Neuroniniai tinklai ir daugiklių analizė, ARIMA, eksponentinis glodinimas, Fourier spektro skaidymas, regresijos ir polinominių vėlinimų analizė, GLM, GLZ, GRM, GDA	Taisyklėmis paremtas ir tikimybinis modeliavimas, neuroniniai tinklai ir kt.	-	-
	Kita		Klasifikacijos ir regresijos medžiai, CHAID metodai, asociacijų taisyklės	Taisyklių indukcija, regresija, asociacijų analizė	Prižiūrimasis modeliavimas (regresija, klasifikacija), neprižiūrimasis modeliavimas (asociacijos, grupavimas)	-
Atvaizdavimas			Pažangios priemonės (3D diagramos, kubo išdėstymo atvaizdavimas, medžių vaizdavimas) ir specializuoti grafiniai įrankiai daugeliui gavybos metodų	Pažangios priemonės (2D, 3D stulpelinės, skritulinės, išsibarstymo diagramos, medžių vaizdavimas ir kt.)	Pažangios priemonės (žemėlapiai, išsibarstymo diagramos, histogramos, sprendimų medžiai ir kt.)	Standartinės priemonės (Stulpelinės diagramos, histogramos)

Atvaizdavimo Internete sprendimų suvestinė palyginant savybes pateikta ***13 – 15 lentelėse***.

13 lentelė. *Atvaizdavimo internete sprendimų lentelė*

Nr.	Analogas
1	IT profesionalų portalas [15]
2	Darbo paieškos sistema „CV.lt“ [24]
3	Klasiokų bendravimo portalas „Klase.lt” [23]
4	IT žinių portalo statistikos modulis

14 lentelė. *Lyginamų savybių lentelė*

Nr.	Savybė
a	Populiarumo sąrašai
b	Laiko intervalo pasirinkimas ir peržiūra laike
c	Bendras objektų pasiskirstymas pagal kiekį
d	Grupuoti duomenys
e	Vidutinės reikšmės
f	Procentinės reikšmės
g	Kritinės reikšmės

15 lentelė. *Analogų savybių palyginimo lentelė*

Nr.	Statistikos peržiūros galimybės						
	a	b	c	d	e	f	g
1	v		v		v	v	v
2	v		v		v	v	
3	v		v		v		
4	v	v	v	v	v	v	

Matome, kad sukurtame statistikos modulio prototipe, lyginant su kitais internetiniais portalų statistikos sprendimais, yra daugiau patogesnių galimybių: laiko pasirinkimo intervalai ir peržiūra laike, bendras objektų pasiskirstymas pagal kiekį, labai svarbūs grupuoti duomenys. O minimali statistika pateikiama daugumoje minėtų internetinių svetainių.

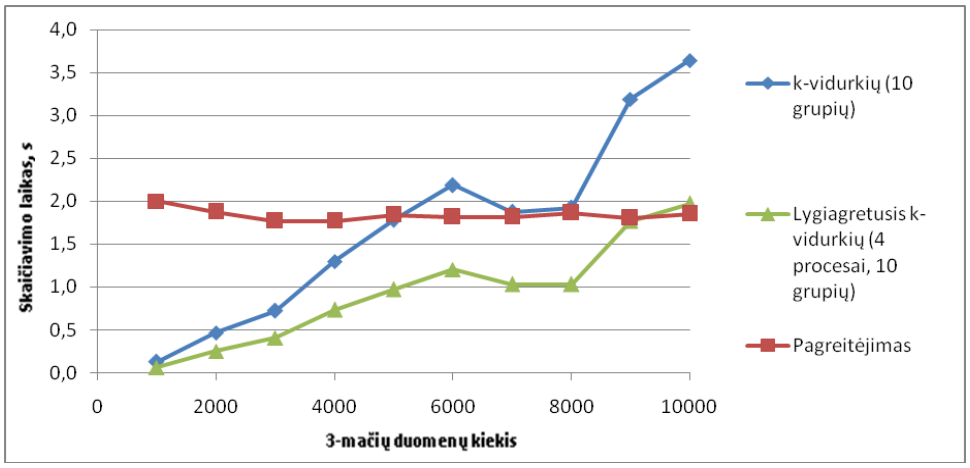
5.2 Duomenų grupavimo eksperimentai

Palyginimui pateikiu tailandiečių pasiūlyto lygiagrečiojo k-vidurkių algoritmo [25], realizuoto žinučių perdavimo sąsajos (MPI) biblioteka C kalboje, kuri naudoja k-vidurkių grupavimo metodą, realizacijos veikimo laiko palyginimą su paprastuoju, kai duomenų labai daug. Eksperimentiškai buvo palyginti grupavimo algoritmai: paprastasis k-vidurkių, lygiagretusis k-vidurkių. Aiškiai savo apdorojimo greičiu išsiskyrė lygiagrečioji k-vidurkių algoritmo realizacija (***16 lentelė***).

16 lentelė. *MPI eksperimentinis tyrimas*

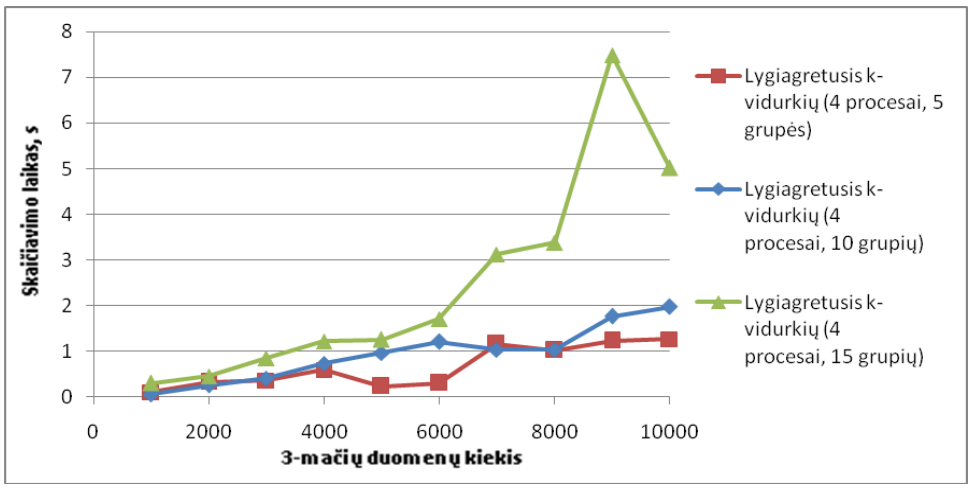
Duomenų kiekis	Paprastasis k-vidurkių alg., *s*	Lygiagretusis globalaus apsikeitimo k-vidurkių alg., *s*	Santykis
5 000 000	341	169	2,01
4 800 000	522	378	1,38
4 500 000	178	188	0,94
4 000 000	148	182	0,81
1 000 000	68	75	0,9

Statistikos modulyje naudojamam grupavimo metodui tirti buvo generuojamos atsitiktinės reikšmės [0 -100] ribose, tas pačias reikšmes grupuojant paprastuoju ir lygiagrečiuoju algoritmu (***27 pav.***). Reikšmių generuota iki dešimties tūkstančių. Matome, kad tokiose duomenų kiekio ribose lygiagretusis algoritmas su 4 lygiagrečiais procesais maždaug du kartus spartesnis už paprastąjį.



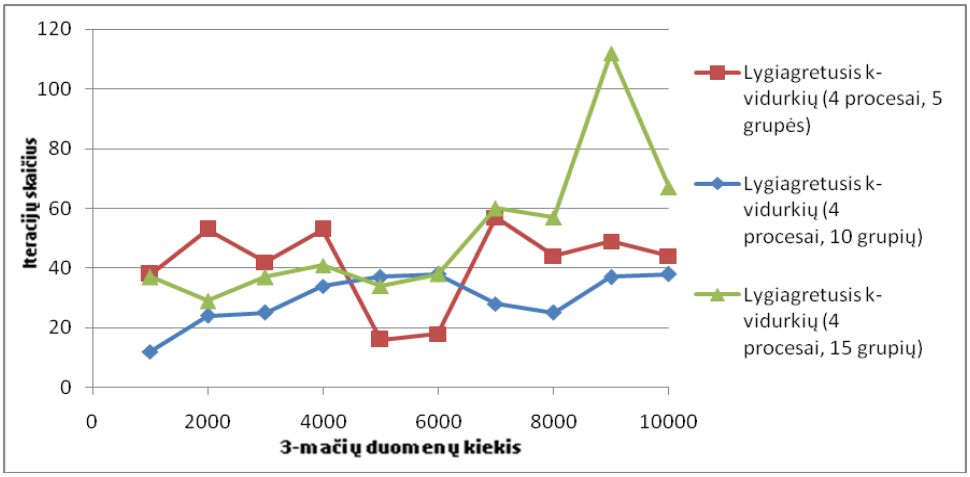
45 pav. *Paprastojo ir lygiagrečiojo k-vidurkių algoritmų skaičiavimo greičio palyginimas*

Taip pat tiriama skaičiavimo greičio priklausomybė nuo grupių kiekio (***46 pav.***). Tie patys duomenys buvo grupuojami į skirtingą kiekį grupių. Rezultate matosi, kad skaičiavimo laikas esant daug grupių (15), sparčiai iššauga.



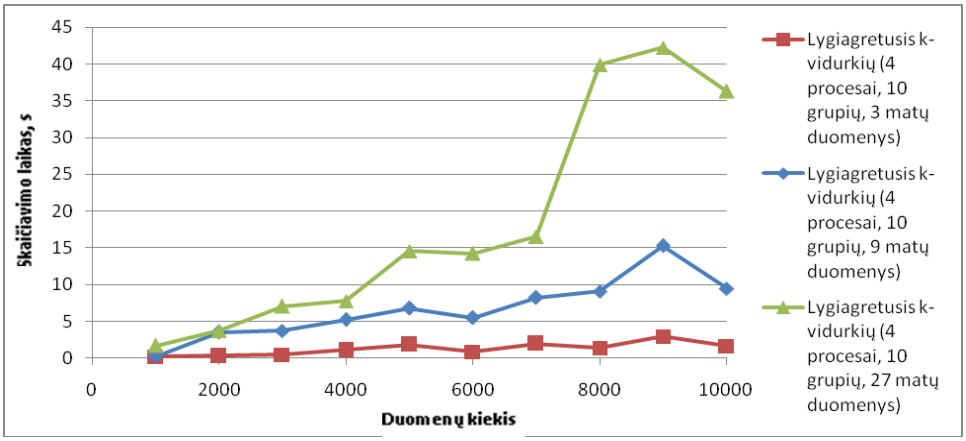
46 pav. *Lygiagrečiojo k-vidurkių algoritmo skaičiavimo greičio palyginimas su skirtingu grupių kiekiu*

Dar algoritmas išbandytas su generuotais skirtingais duomenų rinkiniais, bet vienodu jų kiekiu. Matome, kad skaičiavimo laikas labai susijęs su iteracijų skaičiumi, kuris priklauso nuo duomenų skirtingumo (***47 pav.***).



47 pav. *Lygiagrečiojo k-vidurkių algoritmo iteracijų palyginimas su skirtingu grupių kiekiu*

Augant duomenų matams, stipriai didėja skaičiavimų laikas (***48 pav.***). Jei grupių kiekio padidinimas 3 kartus, skaičiavimo laiką padidino ~8 kartus, tai duomenų matų padidinimas tiek pat kartų, skaičiavimo laiką padidinti gali net 40 kartų.



48 pav. *Lygiagrečiojo k-vidurkių algoritmo skaičiavimo laiko su skirtingais duomenų matais palyginimas*

Matome, kad lygiagretusis algoritmas efektyvesnis už paprastąjį. Bet net lygiagretusis k-vidurkių algoritmas nėra itin tinkamas, kai duomenų kiekiai , duomenų matai, grupių kiekiai labai dideli. Tačiau realizuotam taikymui statistikos modulio prototipe jis tikrai pakankamas, nes duomenų čia nebus milijoniniai kiekiai.

6 IŠVADOS

1. Šiame darbe iškelti tikslai realiam IT žinių portalui IT-EUROPE sukurti eksperimentinio statistikos modulio prototipą buvo pasiekti išanalizavus panašių sistemų modelius ir pasiūlant statistikos modulio prototipą, kurio dalys yra reguliarus duomenų diskretizavimas, duomenų grupavimas su pritaikyta lygiagrečiojo k-vidurkių algoritmo pakoreguota realizacija bei „dotnetcharting“ grafinio atvaizdavimo komponentas.
2. Atlikta statistikos sistemų modelių analizė parodė, kad k-vidurkių metodas yra efektyvus, gerai žinomas ir naudojamas dalies analizuotų sistemų. Jis pritaikomas ne tik statistikos pobūdžio sistemose, bet ir kitose praktinio taikymo srityse (kalbos ar vaizdų atpažinimas, genetikos, geografinių ir kt. duomenų analizė), todėl jį tikrai verta taikyti kuriamai sistemai. O lygiagretųjį k-vidurkių algoritmą, kuris yra spartesnis už paprastąjį, tikslinga taikyti tuomet, kai duomenų yra daug ir juos reikia grupuoti pagal tam tikrą kriterijų ir gauti vizualų rezultatą padedantį priimti tam tikrą sprendimą.
3. Sudarytas koncepcinis modelis artimas duomenų gavybos sistemai. Ją sudaro siūlomas duomenų diskretizavimas turėtų mažiau apsunkinti duomenų bazės darbą nei dirbant su visais portalo sukauptais duomenimis. Parinktas grupavimo metodas būtų duomenų gavybos pagrindas, kurį papildo suprantamas rezultatų atvaizdavimas. Statistikos modulį galima pavadinti DG sistema, nes ji leidžia iš portalo DB sukauptų duomenų išgauti apibendrintą informaciją. Išgautą informaciją pateikus grafiškai, galima interpretuoti duomenis ir stebėti veiklos mastus ar tendencijas.
4. Naudotas komponentų derinys pigus kūrimo kaštų atžvilgiu, kadangi remiamasi patogia pakartotinio panaudojimo technologija. Todėl visų komponentų kurti iš naujo tikrai neapsimoka.
5. Statistikos modulio realizacijai naudota MSSQL Server duomenų bazės valdymo sistema (DBVS) ir pasirinkta Microsoft Visual Studio .Net 2005 programavimo aplinka. ASP.NET(C#) technologija puikiai tinka internetinėms aplikacijoms kurti, o prieigai prie duomenų bazių patogi ADO.NET technologija. Panaudotas MS.NET Framework 2 siūlomas duomenų prieigos lygio atskyrimas.
6. Atlikta veikimo analizė leidžia tikėtis, kad statistikos modulis pasiteisins ir realiomis eksploatavimo sąlygomis: leis lankytojams palyginti vartotojų apkrautumą ir veiksmų aktyvumą, stebėti portalo veiklos mastus ir patogiai bei lanksčiai analizuoti informaciją, kuri administratoriui padės priimti sprendimus dėl sisteminių nustatymų keitimo. Posistemis naudingas tuo, kad nustatytais pjūviais kasdien transformuoja portalo veiklos duomenis ir saugo juos į statistikai skirtas DB lenteles. Sukaupta informacija pateikiama peržiūrai grafiškai. Išskirtinis sistemos bruožas yra galimybė tūkstantinius duomenis norimam periodui grupuoti į norimą kiekį grupių.
7. Darbo rezultatai buvo pristatyti 2007 metų tarpuniversitetinėje doktorantų ir magistrantų konferencijoje „Informacinė visuomenė ir universitetinės studijos“, straipsnis tyrimo tematika išspausdintas konferencijos leidinyje [26] ir pateikiamas *prieduose*.

7 LITERATŪROS ŠALTINIŲ SĄRAŠAS

[1] N. R. Alluri, Evaluation of Data Mining Methods to Support Data Warehouse Administration and Monitoring in SAP Business Warehouse: *magistro tezės*, University of Applied Sciences, Furtwangen, Germany, 2005 balandžio 20, Prieiga per internetą: http://www.3mfuture.com/reiners/thesis_and_diploma/2005_Thesis_Alluri_Data_Mining_Data_Warehouse_SAP-BW.pdf

[2] L. Agosta, The Future of Data Mining – Predictive Analytics, DM Review Magazine, 2004 rugjpūčio numeris

[3] Gartner Group, Data Mining, [žiūrėta 2006-11-30]. Prieiga per internetą: http://www2.nr.no/documents/samba/research_areas/BAMG/Pattern/datamining.html

[4] Wikipedia, Data Mining, [žiūrėta 2007-02-29]. Prieiga per internetą: http://en.wikipedia.org/wiki/Data_mining

[5] K. Thearling, Information About Data Mining and Analytic Technologies, [žiūrėta 2006-11-15]. Prieiga per internetą: http://www.thearling.com/index.htm

[6] R. Šileikienė, Duomenų gavyba – kas tai?, Informacinės technologijos, 2000 m. Leidinio Nr.:11 [žiūrėta 2006-11-11], prieiga per internetą: http://www.it.lt/itweb/it3.nsf/b68106f03be42ba342256cab005d06b2/DFF75F9D8FEEC81842256CAB006639EF/$FILE/informacines_technologijos_11.pdf

[7] Wikipedia, Data Clustering, [žiūrėta 2007-03-01]. Prieiga per internetą: http://en.wikipedia.org/wiki/Data_clustering

[8] P. Berkhin, Survey of Clustering Data Mining Techniques, Accrue Software, Inc., 2002

[9] M. Kantardzic, Data Mining: Concepts, Models, Methods, and Algorithms, FF, 2003

[10] Wikipedia, K-means algorithm, [žiūrėta 2007-03-05]. Prieiga per internetą: http://en.wikipedia.org/wiki/K-means_algorithm#Demonstration_of_the_algorithm

[11] T. Jinlan, Z. Lin, Z. Suqin, L. Lu. Improvement and Parallelism of k-Means Clustering Algorithm, Tsinghua Science and Technology, ISSN 1007-0214 01/21, 277-281 pslp, vol. 10, No. 3, 2005 birželis

[12] T. Kanungo, D. M. Mount, N. S. Netanyahu, Ch. D. Piatko, R. Silverman, A. Y. Wu, An Efficient k-Means Clustering Algorithm: Analysis and Implementation, from IEEE transactions on pattern analysis and machine intelligence, vol. 24, No. 7, 2002 Liepa

[13] SGI, Mineset, [žiūrėta 2007-03-10]. Prieiga per internetą: http://techpubs.sgi.com/library/tpl/cgi-bin/getdoc.cgi/0650/bks/SGI_EndUser/books/MineSetNT_T/sgi_html/pr01.html; http://www.sgi.com/

[14] StatSoft, STATISTICA, [žiūrėta 2007-03-18]. Prieiga per internetą: http://www.statsoft.com/products/products.htm

[15] Rosella, StarProbe, [žiūrėta 2007-03-10]. Prieiga per internetą: http://www.roselladb.com/starprobe.htm

[16] System-admins, Portalas IT sistemų administratoriams, [žiūrėta 2006-11-25]. Prieiga per internetą: http://www.system-admins.net

[17] Interprekyba, Darbo paieškos sistema, [žiūrėta 2006-11-20]. Prieiga per internetą: http://www.cv.lt

[18] MediaWorks, Klasiokų bendravimo portalas, [žiūrėta 2006-11-25]. Prieiga per internetą: http://www.klase.lt

[19] WebAvail Porductions Inc. & Corporate Web Solutions Ltd, DotNetCharting, [žiūrėta 2006-12-25]. Prieiga per internetą: http://www.dotnetcharting.com

[20] J. Gama and C. Pinto, Discretization from Data Streams: Applications to Histograms and Data Mining, LIACC, FEP, University of Porto, 2004

[21] I. Sommerville, Software Engineering, 6th edition, Chapter 14, 2000

[22] Professional Community Server, Free Data Mining Source Code, [žiūrėta 2007-02-25]. Prieiga per internetą: http://www.kdkeys.net/forums/6051/ShowThread.aspx

[23] Microsoft Corporation, Creating a Data Access Layer, [žiūrėta 2007-02-07]. Prieiga per internetą: http://www.asp.net/learn/dataaccess/tutorial01cs.aspx?tabid=63

[24] M. Chapple, Stored Procedures in SQL Server, [žiūrėta 2007-01-19]. Prieiga per internetą: http://databases.about.com/od/sqlserver/l/aastoredprocs.htm

[25] S. Kantabutra, C. Naramittakapong, P. Kornpitak, Pipelined K-means Algorithm on COWs, The Theory of Computation Group, 2003

[26] M. Ruzgys, Portalo statistikos modulis pagrįstas grupavimu: *konferencijos pranešimų medžiaga*. Informacinė visuomenė ir universitetinės studijos [CD-ROM], Vytauto Didžiojo Universitetas, Kaunas, 2007, ISBN 978-9955-12-207-4

8 TERMINŲ IR SANTRUMPŲ ŽODYNĖLIS

Santrumpa, terminas	Paaiškinimas
PĮ	– programinė įranga
DG	– duomenų gavyba
IT (angl. *Information Technology*)	– informacinės technologijos
IS (angl. *Informatikon System*)	– informacijos sistema
DB (angl. *Database)*	– duomenų bazė
SQL (angl. *Structured Query Language*)	– struktūrizuota užklausų kalba
DBVS	– duomenų bazių valdymo sistema
GTrees	– regresijos medžiai
GLM *(*angl. *General Linear Models)*	– bendrieji tiesiniai modeliai
GRM (angl. *General Regression Models*)	– bendrieji regresijos modeliai
GDA (angl. *General Discriminant Analysis*)	– bendroji diskriminantų analizė
PLS (angl. *Partial Least Squares*)	– daline kvadratinė paklaida
EM (angl. *Expectation maximization*)	– tikimybės didinimas
Predictive Analytics	– prognozuojančioji analitika
K-means	– k-vidurkių grupavimo metodas
Machine learning	– save mokančios sistemos
Neural Network	– neuroninis tinklas
Supervised Learning	– prižiūrimasis modeliavimas
Factor Analysis	– daugiklių analizė

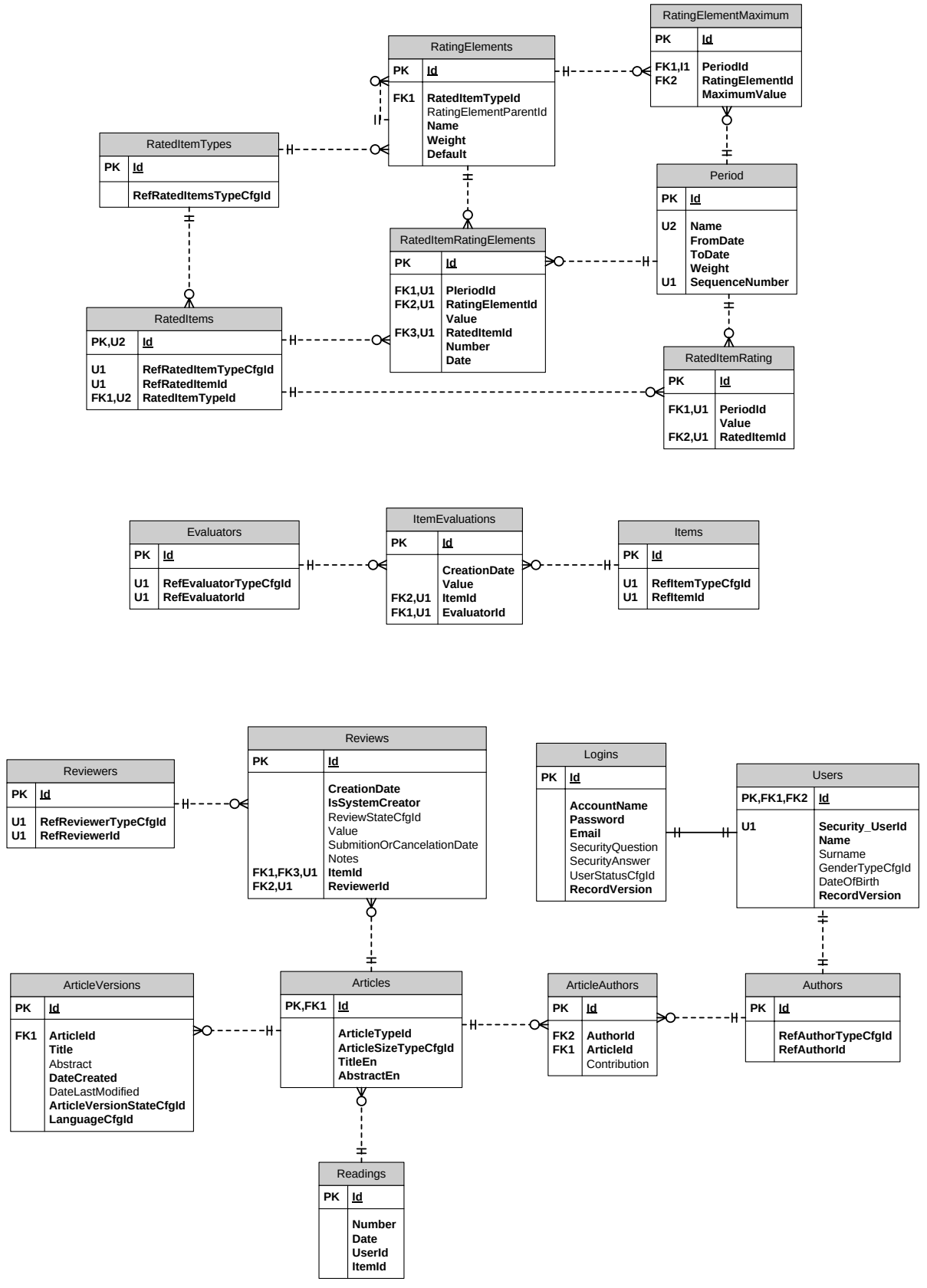
9 PRIEDAI

9.1 Priedas Nr.1 *Serveryje saugomos procedūros*

„getTop10AuthorsCreated“	Išrenka 10 autorių daugiausiai parašiusių straipsnių dabartiniam periode ir kiekiu
„getTop10AuthorsNewCreated“	Išrenka 10 autorių naujausiai parašiusių straipsnį dabartiniam periode ir datą
„getActiveAuthorsTotals“	Išrenka aktyvių autorių, parašiusių bent viena straipsnį, kiekį kasmet
„getAuthorsArticlesOfPeriod“	Išrenka kiekvieno autoriaus straipsnių kiekius per nurodytą periodą
„getPercAuthorArticlesOfPeriod2“	Išrenka nurodyto autoriaus kasmet parašytų straipsnių kiekius, kasmet portale atsiradusių straipsnių kiekius bei procentą, kiek autoriaus straipsniai sudaro portale per metus publikuotų straipsnių kiekio
„getAllAuthorsArticles_avg“	Išrenka kiekvienam periodui straipsnių kiekį, aktyvių autorių kiekį, bei vidutinį straipsnių kiekvienam autoriui per periodą
„getAuthorArticlesForPeriod“	Išrenka nurodyto autoriaus per pasirinktą laiko intervalą parašytų straipsnių kiekius kiekviena
„getPeriods“	Išrenka visus portalo periodus, metus
„getTop10AuthorsRatings“	Išrenka einamojo periodo autorius su aukščiausiu reitingu
„getItemRatingForPeriod“	Išrenka nurodyto autoriaus reitingą laiko intervale. Jį skaičiuoja ir pagal senų periodų reikšmes su svorio koeficientu
„getItemsRatingsOfPeriodEnd“	Išrenka nurodyto periodo pabaigoje autorių/straipsnių reitingus. Juos skaičiuoja ir pagal senų periodų reikšmes su svorio koeficientu
„getAuthorsEvaluationsForPeriod3“	Išrenka autorių vertinimų kiekius per paskutinį periodą
„getAuthorEvaluationsForPeriod“	Išrenka nurodyto autoriaus vertinimų kiekius laiko intervale kasdien
„getTop10Evaluators“	Išrenka per paskutinį periodą 10 daugiausiai vertinusių autorių
„getAuthorsReviewsOfPeriod“	Išrenka nurodyto periodo autorių recenzijų kiekius
„getAuthorReviewsForPeriod“	Išrenka nurodyto laiko intervalo autoriaus recenzijų kiekius
„getReviewsOfPeriod“	Išrenka nurodyto periodo norimo autoriaus recenzijų informaciją: straipsnis, data, būsena
„getArticlesReadingsOfPeriod“	Išrenka nurodyto periodo straipsnių skaitymų kiekius
„getArticleReadingsForPeriod“	Išrenka nurodyto laiko intervalo norimo straipsnio skaitymų kiekius
„getArticleReadingsForPeriod_avg“	Išrenka nurodyto laiko intervalo norimo straipsnio skaitymų kiekį per dieną
„getTop10ItemsReadings“	Išrenka einamojo periodo daugiausiai skaitomus straipsnius ir skaitymų kiekius
„getTotalReadingsPerMonth“	Išrenka einamojo periodo kas mėnesį atliktų skaitymų kiekius
„getTop10NewItems“	Išrenka paskutinių 10 naujausių straipsnių informacija: pavadinimas, data
„getTotalItems“	Išrenka portale sukauptų straipsnių kiekį
„getTotalItemsCreatedPerMonth“	Išrenka portale paskutinį periodą kas mėnesį publikuotų straipsnių kiekį
„getItemRatingForPeriod“	Išrenka nurodyto laiko intervalo norimo straipsnio reitingo kitimą kasdien
„getTop10ArticlesRatings“	Išrenka einamojo periodo pabaigoje 10 straipsnių su aukščiausiu reitingu
„getArticleEvaluationsForPeriod“	Išrenka nurodyto laiko intervalo norimo straipsnio įvertinimų kiekius kiekvienam balui
„getArticleEvaluationsForDate_avg“	Išrenka nurodyto straipsnio paskutinių sukauptų įvertinimų iki nurodytos datos vidurkį
„getArticleEvaluationsForDate“	Išrenka nurodyto straipsnio paskutinius sukauptus įvertinimus iki nurodytos datos
„getArticleEvaluationsForPeriod3a“	Išrenka per paskutinį periodą susikaupusius įvertintų straipsnių kiekius

„getArticleEvaluationsForPeriod3“	Išrenka per paskutinį periodą kasdien įvertintų straipsnių kiekius
„GetarticlesByAuthor“	Išrenka nurodyto autoriaus straipsnius
„getArticlesReviewsOfPeriod“	Išrenka nurodyto periodo straipsnių recenzijų kiekius
„getArticleReviewsForPeriod“	Išrenka nurodyto straipsnio pasirinkto laiko intervalo recenzijų kiekius
„getReviewersOfPeriod“	Išrenka nurodyto periodo autoriaus straipsnių recenzentų informaciją
„getTop10ItemsReviews“	10 paskutinio periodo daugiausiai recenzijų turinčius straipsnius
„getTop10NewReviews“	10 naujausių straipsnių recenzijų

9.2 Priedas Nr.2 *Portalo DB schemos fragmentai*



9.3 Priedas Nr.3 *DB pagrindinių lentelių fragmentai*

„Articles“ lentelės fragmentas:

Id	ArticleId	Title	DateCreated	Articl...	LanguageCfgId
78ddb931-634f-49dd-b45f-23137c81143d	1491a09a-7bdc-43f7-a768-e1bdfaccbc9e	Kodėl spartesnis intern...	2007.01.22...	0	0
8f23e69d-8e2e-4e15-88fc-2eb6c1adfb38	809ac0c7-904b-47ca-949c-16d28de6e150	Programinės įrangos d...	2007.02.09...	0	0
9c543adf-7a87-49b1-aeba-2f61bc8d6816	179450ea-0d26-4165-8735-3051cb99edef	Sonex Sistemos dalyva...	2007.03.08...	0	0
6dcf115e-f118-415c-a1a2-368370ccbea0	8413ba05-d537-4ab1-968e-dd460201daeb	Kas naujo SQL servery...	2007.04.22...	0	0
14600927-8aca-45be-8171-3a5fc31175b9	a2b9358d-3279-4c0f-aa3d-24bfb5567e8f	Programinės įrangos re...	2007.01.01...	0	0
4a6ea68f-36bb-4bb8-9e01-3d41d6bfdfda	5fded706-b774-4530-9a5c-5a09782675bc	Microsoft Office 2007 s...	2007.03.11...	0	0
640558f2-c428-495b-8862-3e6a4e1bff6f	118589ef-e3bf-48e2-8bfe-ed3d5f546d27	Naujausios verslo mod...	2007.02.25...	0	2
36aca17c-cac7-4c57-a055-41d5ae39e1c6	4697a52d-a848-4503-8b86-57c22c7391d0	Verslo planas pagal ele...	2006.12.15...	0	0
2c5a7671-0c74-476c-b89d-45b5b8c514f7	d4417e57-f570-4645-98f0-4ca3d03913b4	Pristatyta nauja „Wind...	2007.01.05...	0	0
0f15b0ca-881a-45e4-8395-46b7117930c7	9a4b1990-773e-4e84-83f2-7c1a76016d1b	UAB "Sonex sistemos" ...	2007.03.15...	0	0
3e56c51b-136d-434e-babc-46c7ff413f85	a11fd8d0-0747-4c37-8c2f-88664f8251d8	„Dell" kompiuteriuose n...	2006.11.12...	0	2
df8e6731-5dff-4ae8-b0ee-49205b628744	af3f1c62-ae5c-43a0-9314-fe0066add50c	„Yahoo" ir „Microsoft" d...	2006.12.20...	0	0
a1069a0d-7146-4742-b9da-52f157b993c9	724d234e-0477-44c5-9cab-b96e61ec1f10	BPĮ ėmėsi iniciatyvos s...	2007.02.15...	0	0
af0fcc34-8505-4cc1-979a-56fbb64fd3e4	4a9778c2-6147-4549-a172-a7893da4eb63	MagicDraw UML 10.5 le...	2007.02.22...	0	0
233a3782-1dad-4f1c-84d7-5b1bf71d2b80	3e63e918-cb9d-4c5d-8759-653798c5ebc0	Microsoft Windows Ser...	2007.04.15...	0	0
0faa733b-ad59-4aa7-b56b-5e34c0e4de04	8dcb8939-f13e-440e-8b46-9e3ea956ba28	"Geležis" muša per IT b...	2006.12.17...	0	1
294d645a-94d4-4415-ae85-60fc1ceee8c3	5fae4e47-58ae-497d-b0ce-877b5262ea56	Po sąstingio IT bendro...	2006.12.16...	0	2
df421483-e4ee-4cdd-971d-6111f4354d22	e840f668-bfc8-44f2-8ce6-c81ba6f57011	Microsoft tikslina verslo...	2007.03.17...	0	0
5e333a67-6695-403d-9680-63ec49f94b08	2ad60fa3-b19d-4fac-bafd-bb3f8eba333a	"PZU Lietuva" investav...	2006.11.19...	0	0
7b8819ef-22f8-4a96-8ccc-68097fb1182e	4fba25fc-368b-4ebd-aa2a-2b446f59af3c	Rusai mano, kad intern...	2006.11.08...	0	1
334d2850-231c-4ee1-be68-6cd80a7b4365	a11fd8d0-0747-4c37-8c2f-88664f8251d6	„Blu-ray" diskų parduot...	2006.11.12...	0	0
c9b75291-e342-436c-b253-6d691364d5e6	716a780b-8f0c-4d58-af4e-f2a1ee445cef	„Microsoft Dynamics N...	2007.03.15...	0	0
1fd184e3-6550-42a0-b84c-715aa2ddb563	a11fd8d0-0747-4c37-8c2f-88664f8251d7	Pagerintas duomenų p...	2006.11.12...	0	1
e05c65a1-84d0-4a03-bf53-76a472e57217	9497df79-f2e0-4cfc-bf36-49e4d2d3c47b	Lietuvoje trūksta geog...	2006.12.15...	0	1
ede821d0-f275-449e-abe8-78b3ab0544ae	eb4b52a6-969b-498b-ad14-a365a9f2d3a4	Verslo planas pagal ele...	2006.11.19...	0	3
2e4da71d-d3d2-4675-960c-7ca4b62061e6	7185cdb9-c4dd-4f6e-8e01-5bf2256f141f	Ketvirtadienį biržoje do...	2007.01.07...	0	0
845c049e-f77c-4f18-a87a-7e4391b6476a	ed7f2e65-b1c1-4621-9584-a52390252898	"Dow Jones" toliau aug...	2007.01.10...	0	0
3ae43d32-2e38-48e9-a731-92e59135865d	a11fd8d0-0747-4c37-8c2f-88664f8251d9	„iPhone" debiutas gali ...	2006.11.13...	0	0
83112f2f-2581-4c3d-8599-94fce6d25b96	91fde060-9063-47b2-a7d0-df1159cc7759	ITG aktyvina veiklą Esti...	2006.11.22...	0	0
2eb13bde-93f7-4a1a-a429-955dd1a19728	61f88a59-e2ea-4357-aabd-557a87ca3f8e	Microsoft planuoja dau...	2007.04.08...	0	0
ebaa053b-7f01-4468-a3c3-98bb011a466f	e30d2bd5-023a-4c31-a1e8-102d07b03950	Vyriausybės svetainė a...	2006.11.05...	0	3
e8574898-672b-48bc-b9d5-9ccf97cd1b94	90ff3541-db43-4227-b6ad-a324c0886f44	Išleista nauja architekt...	2007.02.19...	0	0
41c2e4f9-171c-410b-985c-9de84106fb2e	a53b4320-78cb-4746-a4be-0490d56bfa42	Programinės įrangos ar...	2007.02.05...	0	0
fe4d4c20-f6ca-4a98-a01f-a468f200cf75	2fc650be-657d-46f8-847b-c11fcad031ee	Plinta klaidinantys pran...	2006.12.17...	0	2
1fc4fe93-bb9d-4eeb-87df-a9e19d76ce26	d1e5874a-0d17-4821-851a-61c3fa194ab6	„MicroLink Lietuva" ver...	2006.11.11...	0	2
1d992589-f634-4845-ba86-afbdadaad091	c04c1a79-d544-46ed-84da-96b9cc47fe46	Analitikai rekomenduoj...	2007.01.08...	0	0

Lentelės „Users“ fragmentas:

Id	Security_UserId	Name	Surname	GenderTypeCfgId	DateOfBirth
840eec99-4925-45ee-b7da-04d0f8fced68	b58a4cb1-b942-421e-adef-741fec2b7cee	Ugnė	Ugnaitė	1	1980.01.01 ...
d2773516-40e4-431a-8803-0c3bab664340	073055e2-4203-4f69-a638-e0e6f8655976	Jolanta	Jolantaitė	1	1980.01.01 ...
b52e8bc2-ad88-4c8d-b223-118f7858c138	099baa66-6e8c-4e87-8e5a-c441944b7d88	Milda	Mildaitė	1	1980.01.01 ...
3f8bd8d6-bed6-480f-9b7c-1447765d391a	fbb37846-fc64-4ff3-badf-a48fc166836c	Goda	Godaitytė	1	2000.01.01 ...
ebb6efcb-0b9d-45b1-9c76-1c90e226dd74	80ee71dd-acae-4cca-a60a-f7cfbc071fee	Jonas	Jonaitis	2	1980.01.01 ...
f875e27e-e25e-4c5e-a7a7-262daebc071f	c08d4a63-080e-49ef-936e-9c0ba6d2e09c	Modestas	Modestaitis	2	1980.01.01 ...
fd7c158c-aed2-4e7a-bb68-28b239575a60	ce5ae121-11f0-49ee-ae3b-100865e7b066	Pranas	Pranaitis	2	1982.10.10 ...
4dcce480-f334-4b67-8f83-298dace2ed92	7e07452d-7be9-4c4c-9911-001b4b84092e	Linas	Linaitis	2	1980.01.01 ...
646411f9-4ca3-4e56-855a-392858065cd1	a4fbf5c3-e35f-4a98-bab8-0e60958df858	Saulius	Saulaitis	2	1983.12.15 ...
646411f9-4ca3-4e56-855a-392858065cd5	a4c3ab48-86db-44f9-9bf4-edcd5942b204	Kalvis	Kalvaitis	2	1981.10.01 ...
646411f9-4ca3-4e56-855a-392858065cd6	31dadc0a-0e27-4a67-b01e-74638741e943	Rokas	Rokaitis	2	1975.03.26 ...
2eddc90f-bd4e-46af-9efc-39e060c94be0	2526df35-8ea6-4e06-b422-a3a1ac62030e	Algirdas	Algirdaitis	2	1980.01.01 ...
f927461b-b916-40d9-bd4f-3ab611f2d041	96944201-1d26-4e5e-8ee7-f2b180d547a0	Kęstas	Kęstaitis	2	1980.01.01 ...
702907ae-edfd-4576-9234-3ee42384f008	2c2b8a47-c073-40d3-8278-4cb70d4654cf	Kristina	Kristinaitė	1	1980.01.01 ...
b3a8dbf6-95e7-4853-a231-4864e74293f7	c758aa91-0de5-4b2f-90ca-76925ef655eb	Inga	Ingaitė	1	1980.01.01 ...
d30ae84b-1825-4f35-a1ba-5985d01c727e	cb3adafb-006c-4811-984d-58219ad46710	Marius	Maraitis	2	1980.01.01 ...
3ea4627b-7708-4624-b99b-828ba2b1a866	d3122309-714d-4474-80ab-1df7fae38c96	Markas	Markevičius	2	1963.11.25 ...
4426504c-8741-4df6-bdfd-850ba504577a	0a2785b8-1f41-471b-92f0-28523a3a473c	Jaronimas	Jaraitis	2	1980.01.01 ...
6c61c88d-ff04-4e10-b94b-98bb01217375	45fb8c71-fed7-4d16-9e8c-87297208117c	review-src	surreview-src	0	2007.01.19 ...
5e132bca-0a1b-4fa5-aa52-98bb0121751b	f3ab5bcf-42f0-4884-9c10-482b34cdd3c9	review-dst0	surreview-dst0	0	2007.01.19 ...
6c5c0200-92e1-471c-915d-98bb01217696	3dc540ed-bd74-48d9-b95c-68b003b998e5	review-dst1	surreview-dst1	0	2007.01.19 ...
16644a3e-5f6b-4dcc-8ef8-9d8d7c0508d5	65156404-321b-4e8c-a1f0-8028dca7bfdb	Petras	Petraitis	2	1980.01.01 ...
cc6e4984-4681-414c-9180-9fc136f03b84	8bd09d16-9c24-4d6a-b1f2-e89dd9693749	Kazys	Kazaitis	2	1980.01.01 ...
7ada731e-2c6d-4c6e-8610-bba66cfbc15e	4515cb19-6a9f-44f7-907a-4cd6a7f6150e	Vaidotas	Vaidotaitis	2	1980.01.01 ...
88560fb9-0b44-4a14-9d86-c1732b17c402	7364d54f-4ed6-42c4-9c0b-8c520696fe1e	John	Johnson	2	1999.01.06 ...
724de820-491e-47f4-9bdb-d238aa64229f	1f2aa187-8118-4561-a5ae-ae99e5ada7f2	Radvila	Radvilaitė	1	1980.01.01 ...
8d686e82-028b-4730-b6b2-d5aa5e6ea44b	94b6b4a0-4641-482c-aa48-61c6d24cd956	Gintaras	Gintaraitis	2	1980.01.01 ...
05479bb4-e56b-48a3-8f94-d63de20d6d35	a2494919-abc5-4e83-a10d-345694f32dd5	Leonas	Leonaitis	2	1955.11.08 ...
c1d24693-15c5-4840-9b66-de4cda83387b	ce686d9c-f1cf-4fd7-b96e-67e5140bbad4	Vytautas	Vytautaitis	2	1980.01.01 ...
bd3ac6a0-e0c2-422d-84d2-e7a7932bef70	57cfe570-073e-4af9-bbcd-6b3a9b0eafbc	Rūta	Rūtaitė	1	1980.01.01 ...
5c628f6e-7aa3-4669-8433-f409ce64053f	ba79c4b0-9c7c-410a-8c29-628b24043fe4	Martynas	Martynaitis	2	1980.01.01 ...

Lentelės „Authors“ fragmentas:

Id	RefAuthorTypeCfgId	RefAuthorId
73db6825-96c1-4a84-8567-04ab3e34d96c	0	8d686e82-028b-4730-b6b2-d5aa5e6ea44b
3adf140a-c89d-4c85-a46c-065d0917e8e8	0	840eec99-4925-45ee-b7da-04d0f8fced68
edc0ca2a-21a0-4fa2-be5d-0cf6195098ee	0	c1d24693-15c5-4840-9b66-de4cda83387b
f72dac4e-4a34-4d1f-a376-178eb8895d62	0	f875e27e-e25e-4c5e-a7a7-262daebc071f
ca0a2d6f-aa19-4337-b10c-39a87df28066	0	d2773516-40e4-431a-8803-0c3bab664340
71da1e27-fe22-4b70-b20f-4af727019f47	0	b52e8bc2-ad88-4c8d-b223-118f7858c138
12e73d42-1885-4899-a9c0-5758923c5a08	0	05479bb4-e56b-48a3-8f94-d63de20d6d35
9d5fd66e-6989-48f2-9dce-5d3c1830b681	0	3f8bd8d6-bed6-480f-9b7c-1447765d391a
0db8dfae-83e4-4fda-8c15-5fe0f98a036b	0	ebb6efcb-0b9d-45b1-9c76-1c90e226dd74
7baebb11-b05c-4d89-8680-621031486d53	0	f875e27e-e25e-4c5e-a7a7-262daebc071f
d9ec0170-e162-4d47-8609-65cc2949ebb5	0	fd7c158c-aed2-4e7a-bb68-28b239575a60
aff7ece6-5879-410a-9b3c-6f3c9636e9c7	0	4dcce480-f334-4b67-8f83-298dace2ed92
8f5322f6-59ae-427a-8417-73696f713a68	0	646411f9-4ca3-4e56-855a-392858065cd1
ed5d391c-d605-4fda-8291-75b5732ae9fe	0	646411f9-4ca3-4e56-855a-392858065cd5
0adbe4ae-0b78-43a7-b6f5-7746a817c5cb	0	646411f9-4ca3-4e56-855a-392858065cd6
fdbd969e-3996-497f-8374-81f6d6d2997d	0	2eddc90f-bd4e-46af-9efc-39e060c94be0
cc14bed7-6438-4ba6-8231-829729afd3db	0	f927461b-b916-40d9-bd4f-3ab611f2d041
7b881bef-8650-4841-8482-95705747d9da	0	702907ae-edfd-4576-9234-3ee42384f008
067ba3fa-82b8-4639-be67-9657890bef9b	0	724de820-491e-47f4-9bdb-d238aa64229f
1f808b1c-33b7-446c-a0e1-98bb011a4661	0	b3a8dbf6-95e7-4853-a231-4864e74293f7
1a61bce3-5c59-4d13-a41a-a6784c0cca20	0	5c628f6e-7aa3-4669-8433-f409ce64053f
e0a3f759-ae22-41ac-b665-aa8eaa64b581	0	d30ae84b-1825-4f35-a1ba-5985d01c727e
9afd9653-9f49-4d72-8aa1-ad57383aecc2	0	3ea4627b-7708-4624-b99b-828ba2b1a866
d8d3d1e8-41af-42b3-9cc7-b58b28406798	0	bd3ac6a0-e0c2-422d-84d2-e7a7932bef70
63a53dff-0849-4592-9e94-ccc2074280a2	0	88560fb9-0b44-4a14-9d86-c1732b17c402
0ffa2832-a8bd-441a-8f9e-cf374e82b73c	0	4426504c-8741-4df6-bdfd-850ba504577a
e8bbf350-b182-41d7-bb48-e0d120fc1c76	0	16644a3e-5f6b-4dcc-8ef8-9d8d7c0508d5
97e359ca-37b9-4304-91e1-eb932853683c	0	7ada731e-2c6d-4c6e-8610-bba66cfbc15e
0b5d62e4-2543-4db1-a400-fa11c585c865	0	cc6e4984-4681-414c-9180-9fc136f03b84

Lentelės „ArticlesAuthors“ fragmentas:

Id	AuthorId	ArticleId	Contribution	IsPrime
36d57391-d4a3-4f75-b7cc-0b0824b4ac93	97e359ca-37b9-4304-91e1-eb932853683c	cffa573f-55de-4ff1-b201-ee68d44aa0ed	0,3	False
e68750e2-6f7b-4d46-9d73-0d7301d24abb	73db6825-96c1-4a84-8567-04ab3e34d96c	fafeead0-5905-42f7-9dab-a21ede6d9787	0,2	False
2811fdf4-bc9b-4636-898b-0e9b94b78867	3adf140a-c89d-4c85-a46c-065d0917e8e8	e30d2bd5-023a-4c31-a1e8-102d07b03950	1	False
91bf9531-984d-498e-8168-0f3897191b61	edc0ca2a-21a0-4fa2-be5d-0cf6195098ee	fafeead0-5905-42f7-9dab-a21ede6d9787	0,2	False
e546d7ee-f598-4d79-91cb-0ff757d6f99e	e0a3f759-ae22-41ac-b665-aa8eaa64b581	1491a09a-7bdc-43f7-a768-e1bdfaccbc9e	1	False
1871fde0-3f9d-4bb5-9a02-14cb7f8d5f7b	f72dac4e-4a34-4d1f-a376-178eb8895d62	63f6f39e-f14c-4152-b34d-94deb99a9dbb	0,5	False
ce05cf8d-ac3b-4fa6-89a4-17ac4a766056	1f808b1c-33b7-446c-a0e1-98bb011a4661	cffa573f-55de-4ff1-b201-ee68d44aa0ed	0,3	False
f72842f8-3ef9-42e7-bc05-183d3ad49178	aff7ece6-5879-410a-9b3c-6f3c9636e9c7	8dcb8939-f13e-440e-8b46-9e3ea956ba28	1	False
ace8caf5-ae32-46b3-a62e-1cfdd2d37d7c	0b5d62e4-2543-4db1-a400-fa11c585c865	a1441359-9430-45b6-b590-eaebd8e2e982	1	False
4038831f-8ef0-486b-b9a8-20e9810e93a8	1a61bce3-5c59-4d13-a41a-a6784c0cca20	2fc650be-657d-46f8-847b-c11fcad031ee	1	False
c37d6d86-2945-4c1b-8c73-2161f82ce356	9afd9653-9f49-4d72-8aa1-ad57383aecc2	8413ba05-d537-4ab1-968e-dd460201daeb	1	False
89f67702-9637-44c9-b942-24c4f1bf5f16	cc14bed7-6438-4ba6-8231-829729afd3db	809ac0c7-904b-47ca-949c-16d28de6e150	1	False
df52f9b7-17f8-4372-8e5e-25fe993f5f6d	0adbe4ae-0b78-43a7-b6f5-7746a817c5cb	179450ea-0d26-4165-8735-3051cb99edef	1	False
7663edc4-c617-4af6-b578-26f3dd9b15ed	1f808b1c-33b7-446c-a0e1-98bb011a4661	6d5a320d-da6f-4aca-8497-76f6de1f35f6	1	False
50d74505-3aac-4c49-9560-274918cc9b81	ca0a2d6f-aa19-4337-b10c-39a87df28066	07f8c69e-eed9-4871-89af-91fcc2be6909	0,3	False
f3f68285-8348-4fe1-80d1-2a83600957c8	d8d3d1e8-41af-42b3-9cc7-b58b28406798	f60f431c-4198-427f-a1dc-aa8ac6095ef1	1	False
38a49467-f882-4373-a889-2ad7d2f967ff	71da1e27-fe22-4b70-b20f-4af727019f47	4abffd1b-a125-4fb3-9737-12a37ee26714	0,6	False
d78de511-9e0f-449a-a8d0-2b5356d24539	12e73d42-1885-4899-a9c0-5758923c5a08	91fde060-9063-47b2-a7d0-df1159cc7759	1	False
741ca161-9e23-44dc-afc2-2c347ae820d8	71da1e27-fe22-4b70-b20f-4af727019f47	d619b419-6bbf-4c3c-83d3-e53cb5109acc	1	False
bd7d2b71-7e63-4d35-ae96-3859336173 21	ca0a2d6f-aa19-4337-b10c-39a87df28066	41aff43c-cbb6-48b7-b106-03aa77642324	1	False
e7ffc712-63ba-4300-aa60-394e6c289c48	9d5fd66e-6989-48f2-9dce-5d3c1830b681	8511c03f-db20-4c8b-a532-ab0821ea4b08	1	False
b62460dd-0b34-4dee-a9ab-3db9a954d733	0db8dfae-83e4-4fda-8c15-5fe0f98a036b	fafeead0-5905-42f7-9dab-a21ede6d9787	0,6	False
0e6b1e17-a054-468e-849e-3f0fff7085ef	0b5d62e4-2543-4db1-a400-fa11c585c865	2ad60fa3-b19d-4fac-bafd-bb3f8eba333a	1	False
16a7cdb8-583b-42b4-848e-41748ef1bb53	067ba3fa-82b8-4639-be67-9657890bef9b	61f88a59-e2ea-4357-aabd-557a87ca3f8e	1	False
0848b3a3-1343-4202-8a1b-4345299a75be	7baebb11-b05c-4d89-8680-621031486d53	eb4b52a6-969b-498b-ad14-a365a9f2d3a4	1	False
9a940bcb-c367-47b1-8134-456eb885b1e1	d9ec0170-e162-4d47-8609-65cc2949ebb5	0f3f8d62-4083-4386-9cd9-194453ebfe4d	0,3	False
eaaf9579-ec0a-4712-b403-4609bcfe9f32	aff7ece6-5879-410a-9b3c-6f3c9636e9c7	f00820e1-fa11-4511-9804-5962d8c952d9	1	False
036ad199-0016-4254-bc8c-4ac8245dabd1	63a53dff-0849-4592-9e94-ccc2074280a2	0f3f8d62-4083-4386-9cd9-194453ebfe4d	0,3	False
a1dc2253-bdae-4ac6-a29b-4b0a8fd82dd6	9d5fd66e-6989-48f2-9dce-5d3c1830b681	af3f1c62-ae5c-43a0-9314-fe0066add50c	0,5	False
3122b1b2-2716-46df-b6e8-5147c5dd96c5	63a53dff-0849-4592-9e94-ccc2074280a2	0aa29327-f14d-4e75-b82d-e33be1931eff	0,3	False
5f9dd318-d635-44b3-b8da-529044fafeb5	97e359ca-37b9-4304-91e1-eb932853683c	8d4753ab-897a-45d7-bb26-1f15933f900a	1	False
2dcecd87-5bff-4d9c-ba48-60a986762902	ed5d391c-d605-4fda-8291-75b5732ae9fe	0aa29327-f14d-4e75-b82d-e33be1931eff	0,3	False
ea49d074-2aa5-476d-834c-64a976dd5f92	12e73d42-1885-4899-a9c0-5758923c5a08	724d234e-0477-44c5-9cab-b96e61ec1f10	1	False
152b3bf2-8300-4d99-9f45-68a045f99a7b	e8bbf350-b182-41d7-bb48-e0d120fc1c76	7185cdb9-c4dd-4f6e-8e01-5bf2256f141f	0,5	False
dc1b9fb7-b508-4216-b78c-6ffa05574f0e	9afd9653-9f49-4d72-8aa1-ad57383aecc2	0aa29327-f14d-4e75-b82d-e33be1931eff	0,3	False
08e466a4-4a13-4d55-85e1-73eaece4d926	8f5322f6-59ae-427a-8417-73696f713a68	d1e5874a-0d17-4821-851a-61c3fa194ab6	1	False

Lentelės „Articles_Stat“ fragmentas:

Date	Lng	AuthorId	Cnt
2006.11.22 ...	0	12e73d42-1885-...	1
2006.12.15 ...	0	edc0ca2a-21a0-...	1
2006.12.15 ...	1	1f808b1c-33b7-...	1
2006.12.16 ...	0	d8d3d1e8-41af-...	1
2006.12.16 ...	1	0ffa2832-a8bd-...	1
2006.12.16 ...	2	cc14bed7-6438-...	1
2006.12.17 ...	1	aff7ece6-5879-...	1
2006.12.17 ...	2	1a61bce3-5c59-...	1
2006.12.20 ...	0	9d5fd66e-6989-...	1
2006.12.20 ...	0	aff7ece6-5879-...	1
2007.01.01 ...	0	f72dac4e-4a34-...	1
2007.01.02 ...	0	d9ec0170-e162-...	1
2007.01.05 ...	0	d8d3d1e8-41af-...	1
2007.01.07 ...	0	71da1e27-fe22-...	1
2007.01.07 ...	0	e8bbf350-b182-...	1
2007.01.08 ...	0	0db8dfae-83e4-...	1
2007.01.08 ...	0	1a61bce3-5c59-...	1
2007.01.10 ...	0	7b881bef-8650-...	1
2007.01.18 ...	0	d8d3d1e8-41af-...	1
2007.01.19 ...	0	7baebb11-b05c-...	1
2007.01.19 ...	0	cc14bed7-6438-...	1
2007.01.19 ...	0	1a61bce3-5c59-...	1
2007.01.22 ...	0	ed5d391c-d605-...	1
2007.01.22 ...	0	e0a3f759-ae22-...	1
2007.01.22 ...	0	9afd9653-9f49-...	1
2007.01.22 ...	0	63a53dff-0849-...	1
2007.01.23 ...	0	0b5d62e4-2543-...	1
2007.02.02 ...	0	ca0a2d6f-aa19-...	1
2007.02.05 ...	0	8f5322f6-59ae-...	1
2007.02.09 ...	0	cc14bed7-6438-...	1
2007.02.15 ...	0	12e73d42-1885-...	1
2007.02.17 ...	0	1a61bce3-5c59-...	1
2007.02.19 ...	0	3adf140a-c89d-...	1
2007.02.22 ...	0	f72dac4e-4a34-...	1
2007.02.25 ...	2	aff7ece6-5879-...	1
2007.03.02 ...	0	7baebb11-b05c-...	1

Lentelės „ItemEvaluations_Stat_1“ fragmentas:

Date	Value	Cnt	ItemId
2007.03.28...	0,1	2	2fc650be-657d-46f8-847b-c11fcad031ee
2007.03.28...	0,2	1	809ac0c7-904b-47ca-949c-16d28de6e150
2007.03.28...	0,2	1	9a4b1990-773e-4e84-83f2-7c1a76016d1b
2007.03.28...	0,3	2	f60f431c-4198-427f-a1dc-aa8ac6095ef1
2007.03.28...	0,4	2	53e71344-66a6-458b-937f-6b598bc507d6
2007.03.28...	0,5	2	e30d2bd5-023a-4c31-a1e8-102d07b03950
2007.03.29...	0,5	1	f60f431c-4198-427f-a1dc-aa8ac6095ef1
2007.03.29...	0,5	2	1491a09a-7bdc-43f7-a768-e1bdfaccbc9e
2007.04.10...	0,2	3	af3f1c62-ae5c-43a0-9314-fe0066add50c
2007.04.11...	0,2	10	41aff43c-cbb6-48b7-b106-03aa77642324
2007.04.11...	0,2	1	d1e5874a-0d17-4821-851a-61c3fa194ab6
2007.04.11...	0,2	3	5aae7bdf-5d40-4051-8442-9e6efb6b917b
2007.04.11...	0,2	8	2ad60fa3-b19d-4fac-bafd-bb3f8eba333a
2007.04.11...	0,3	1	d1e5874a-0d17-4821-851a-61c3fa194ab6
2007.04.11...	0,4	10	d1e5874a-0d17-4821-851a-61c3fa194ab6
2007.04.12...	0,2	4	c04c1a79-d544-46ed-84da-96b9cc47fe46
2007.04.12...	0,4	5	e6b890cc-c03c-48c9-9f74-5ceb43db4e4d
2007.04.13...	0,1	6	d1e5874a-0d17-4821-851a-61c3fa194ab6
2007.04.13...	0,2	3	118589ef-e3bf-48e2-8bfe-ed3d5f546d27
2007.04.13...	0,3	1	d1e5874a-0d17-4821-851a-61c3fa194ab6
2007.04.13...	0,4	4	8511c03f-db20-4c8b-a532-ab0821ea4b08
2007.04.13...	0,5	1	d1e5874a-0d17-4821-851a-61c3fa194ab6
2007.04.14...	0,2	7	2fc650be-657d-46f8-847b-c11fcad031ee
2007.04.14...	0,4	4	f3542ee5-92dc-41ef-a5d8-2aa5e6f64432
2007.04.15...	0,3	2	4697a52d-a848-4503-8b86-57c22c7391d0
2007.04.15...	0,4	2	d4417e57-f570-4645-98f0-4ca3d03913b4
2007.04.16...	0,3	10	a11fd8d0-0747-4c37-8c2f-88664f8251d8
2007.04.16...	0,5	12	af3f1c62-ae5c-43a0-9314-fe0066add50c
2007.04.17...	0,3	3	41aff43c-cbb6-48b7-b106-03aa77642324
2007.04.17...	0,3	2	809ac0c7-904b-47ca-949c-16d28de6e150
2007.04.17...	0,4	3	6e5f480f-88f3-446b-aae9-36ecd6f10232
2007.04.17...	0,4	2	724d234e-0477-44c5-9cab-b96e61ec1f10
2007.04.17...	0,5	4	9497df79-f2e0-4cfc-bf36-49e4d2d3c47b
2007.04.17...	0,5	2	7185cdb9-c4dd-4f6e-8e01-5bf2256f141f
2007.04.17...	0,5	3	90ff3541-db43-4227-b6ad-a324c0886f44
2007.04.18...	0,1	4	fafeead0-5905-42f7-9dab-a21ede6d9787

Lentelės „Readings_Stat“ fragmentas:

Date	ItemId	Cnt
2007.02.27 ...	809ac0c7-904b-47ca-949c-16d28de6e150	1
2007.02.28 ...	4697a52d-a848-4503-8b86-57c22c7391d0	1
2007.02.28 ...	e6b890cc-c03c-48c9-9f74-5ceb43db4e4d	1
2007.03.01 ...	5fae4e47-58ae-497d-b0ce-877b5262ea56	1
2007.03.02 ...	6e5f480f-88f3-446b-aae9-36ecd6f10232	1
2007.03.03 ...	90ff3541-db43-4227-b6ad-a324c0886f44	1
2007.03.04 ...	a53b4320-78cb-4746-a4be-0490d56bfa42	1
2007.03.05 ...	179450ea-0d26-4165-8735-3051cb99edef	1
2007.03.05 ...	5fded706-b774-4530-9a5c-5a09782675bc	1
2007.03.05 ...	eb4b52a6-969b-498b-ad14-a365a9f2d3a4	1
2007.03.05 ...	118589ef-e3bf-48e2-8bfe-ed3d5f546d27	1
2007.03.06 ...	e30d2bd5-023a-4c31-a1e8-102d07b03950	2
2007.03.06 ...	0f3f8d62-4083-4386-9cd9-194453ebfe4d	3
2007.03.06 ...	f3542ee5-92dc-41ef-a5d8-2aa5e6f64432	11
2007.03.06 ...	d1e5874a-0d17-4821-851a-61c3fa194ab6	1
2007.03.06 ...	a11fd8d0-0747-4c37-8c2f-88664f8251d6	2
2007.03.06 ...	a11fd8d0-0747-4c37-8c2f-88664f8251d8	9
2007.03.06 ...	63f6f39e-f14c-4152-b34d-94deb99a9dbb	4
2007.03.06 ...	fcd045d7-3c89-4825-8ed2-98bb011a4665	3
2007.03.06 ...	eb4b52a6-969b-498b-ad14-a365a9f2d3a4	8
2007.03.06 ...	2ad60fa3-b19d-4fac-bafd-bb3f8eba333a	1
2007.03.06 ...	91fde060-9063-47b2-a7d0-df1159cc7759	1
2007.03.07 ...	e30d2bd5-023a-4c31-a1e8-102d07b03950	4
2007.03.07 ...	f3542ee5-92dc-41ef-a5d8-2aa5e6f64432	6
2007.03.07 ...	a11fd8d0-0747-4c37-8c2f-88664f8251d6	3
2007.03.07 ...	a11fd8d0-0747-4c37-8c2f-88664f8251d8	6
2007.03.07 ...	63f6f39e-f14c-4152-b34d-94deb99a9dbb	5
2007.03.07 ...	90ff3541-db43-4227-b6ad-a324c0886f44	1
2007.03.08 ...	c06d6518-a175-406f-b323-0b1cc114195c	1
2007.03.08 ...	0f3f8d62-4083-4386-9cd9-194453ebfe4d	4
2007.03.08 ...	53e71344-66a6-458b-937f-6b598bc507d6	1
2007.03.08 ...	a11fd8d0-0747-4c37-8c2f-88664f8251d6	1
2007.03.08 ...	a11fd8d0-0747-4c37-8c2f-88664f8251d8	2
2007.03.08 ...	c04c1a79-d544-46ed-84da-96b9cc47fe46	2
2007.03.08 ...	eb4b52a6-969b-498b-ad14-a365a9f2d3a4	1
2007.03.09 ...	a53b4320-78cb-4746-a4be-0490d56bfa42	1

Lentelės „Item_Reviews_Stat“ fragmentas:

Date	ItemId	Cnt
2006.12.11...	9497df79-f2e0-4cfc-bf36-49e4d2d3c47b	1
2006.12.15...	63f6f39e-f14c-4152-b34d-94deb99a9dbb	1
2006.12.17...	e6b890cc-c03c-48c9-9f74-5ceb43db4e4d	1
2006.12.18...	19599b3b-0ddc-45bf-a792-7baf9e37212a	1
2006.12.18...	a11fd8d0-0747-4c37-8c2f-88664f8251d6	1
2006.12.20...	8dcb8939-f13e-440e-8b46-9e3ea956ba28	1
2006.12.29...	af3f1c62-ae5c-43a0-9314-fe0066add50c	1
2007.01.04...	e6b890cc-c03c-48c9-9f74-5ceb43db4e4d	1
2007.01.06...	91fde060-9063-47b2-a7d0-df1159cc7759	1
2007.01.08...	9497df79-f2e0-4cfc-bf36-49e4d2d3c47b	1
2007.01.11...	f00820e1-fa11-4511-9804-5962d8c952d9	1
2007.01.21...	ed7f2e65-b1c1-4621-9584-a52390252898	1
2007.01.22...	d4417e57-f570-4645-98f0-4ca3d03913b4	1
2007.01.27...	f60f431c-4198-427f-a1dc-aa8ac6095ef1	1
2007.01.27...	498ab14e-4e0d-4b3d-8768-b0516421293e	1
2007.01.29...	0aa29327-f14d-4e75-b82d-e33be1931eff	1
2007.01.29...	a1441359-9430-45b6-b590-eaebd8e2e982	1
2007.02.13...	fafeead0-5905-42f7-9dab-a21ede6d9787	1
2007.02.18...	a53b4320-78cb-4746-a4be-0490d56bfa42	1
2007.02.18...	4abffd1b-a125-4fb3-9737-12a37ee26714	1
2007.02.19...	6e5f480f-88f3-446b-aae9-36ecd6f10232	1
2007.02.20...	d9ec0170-e162-4d47-8609-65cc2949ebb5	1
2007.02.20...	90ff3541-db43-4227-b6ad-a324c0886f44	1
2007.02.21...	809ac0c7-904b-47ca-949c-16d28de6e150	1
2007.02.21...	724d234e-0477-44c5-9cab-b96e61ec1f10	1
2007.02.22...	41aff43c-cbb6-48b7-b106-03aa77642324	1
2007.02.25...	4a9778c2-6147-4549-a172-a7893da4eb63	1
2007.02.28...	19599b3b-0ddc-45bf-a792-7baf9e37212a	1
2007.02.28...	118589ef-e3bf-48e2-8bfe-ed3d5f546d27	1
2007.03.05...	9a4b1990-773e-4e84-83f2-7c1a76016d1b	1
2007.03.07...	e840f668-bfc8-44f2-8ce6-c81ba6f57011	1
2007.03.08...	a11fd8d0-0747-4c37-8c2f-88664f8251d6	1
2007.03.11...	809ac0c7-904b-47ca-949c-16d28de6e150	1
2007.03.11...	5fded706-b774-4530-9a5c-5a09782675bc	1
2007.03.13...	fafeead0-5905-42f7-9dab-a21ede6d9787	1
2007.03.15...	c06d6518-a175-406f-b323-0b1cc114195c	1

Lentelės „RatedItemRating_Stat“ fragmentas:

Date	PeriodId	Value	RatedItemId
2007.01.13 ...	69896c7d-e68e-...	0,0150000006	e1732dd4-b4d6-461b-9934-393cc087f2fa
2007.01.13 ...	69896c7d-e68e-...	0,024	b5152b0e-8ec4-495e-a871-3aaeb7ad6c5c
2007.01.13 ...	69896c7d-e68e-...	0,0222000014	ba538dc1-b1bf-4e93-950e-53dad065d4e8
2007.01.13 ...	69896c7d-e68e-...	0,0336	3b9540a8-685b-451c-ac0b-599c258bf772
2007.01.13 ...	69896c7d-e68e-...	0,016	c7f47b93-92cf-45fc-b2e9-5f85cf495a08
2007.01.13 ...	69896c7d-e68e-...	0,0100000007	bef13210-55d6-4c26-ba52-5f86aac17939
2007.01.13 ...	69896c7d-e68e-...	0,021	c5dbcb4a-cebf-4d7f-92cc-6358e01f19d3
2007.01.13 ...	69896c7d-e68e-...	0,011	941bf566-d7dc-41f4-8124-63605515955e
2007.01.13 ...	69896c7d-e68e-...	0,0175	bc4ea338-bc15-4189-b83f-655236462df7
2007.01.13 ...	69896c7d-e68e-...	0,007	13935674-2ecc-4260-8a4c-671f38c65bfd
2007.01.13 ...	69896c7d-e68e-...	0,0139999995	4538e9c0-2cc5-4aef-a43d-67ab4416fbe4
2007.01.13 ...	69896c7d-e68e-...	0,001	bc75e4f1-d2dc-482c-9e1b-6a5f121a599a
2007.01.13 ...	69896c7d-e68e-...	0,001	8d5f8d89-951c-4579-a893-6b0e3a09c999
2007.01.13 ...	69896c7d-e68e-...	0,001	caf323eb-1add-4c35-911b-72eb5d64a91e
2007.01.13 ...	69896c7d-e68e-...	0,022	4d890619-b7ef-4146-b3b6-7490dd563a18
2007.01.13 ...	69896c7d-e68e-...	0,001	2dd269ce-b832-4db5-bfcb-771ccb8bb9ba
2007.01.13 ...	69896c7d-e68e-...	0,024	1fc48901-6bb4-42d0-b982-78677f9509d0
2007.01.13 ...	69896c7d-e68e-...	0,0409999974	b7272b6e-e45d-4d79-8b51-798de294a7a7
2007.01.13 ...	69896c7d-e68e-...	0,001	cc30175c-3690-4921-bc0e-7b022abeec2b
2007.01.13 ...	69896c7d-e68e-...	0,0200000014	1789af44-2ce4-4a45-a1b4-7fa9e5a653bd
2007.01.13 ...	69896c7d-e68e-...	0,024	08515e82-ecd7-4289-aa6f-83f7a96544b7
2007.01.13 ...	69896c7d-e68e-...	0,0302999988	9f6c0f99-529e-4499-8306-85a92fcdfca7
2007.01.13 ...	69896c7d-e68e-...	0,0200000014	a683faa6-cf8e-4ae4-92cd-8b42ae4ff5ec
2007.01.13 ...	69896c7d-e68e-...	0,0574000031	eb8c00b3-f859-4de1-aa30-8d0a75f4a202
2007.01.13 ...	69896c7d-e68e-...	0,011	58534e64-f022-48ae-9b93-9471e9309179
2007.01.13 ...	69896c7d-e68e-...	0,001	268c95b7-cbf6-4661-bbba-9e30ec165158
2007.01.13 ...	69896c7d-e68e-...	0,0175	dddb241b-b77b-4fd1-be71-a1f5efb4be47
2007.01.13 ...	69896c7d-e68e-...	0,0269999988	8d682433-32e6-475e-bb49-a7db9ea9f84f
2007.01.13 ...	69896c7d-e68e-...	0,022499999	1292b47b-c2c1-465d-8ca2-ae0f37561ba1
2007.01.13 ...	69896c7d-e68e-...	0,021	9852ac27-8d07-4d84-8225-bc3ea8f607b7
2007.01.13 ...	69896c7d-e68e-...	0,0117000006	c4ca91c5-ba43-4465-a1c5-c5da74a76dbe
2007.01.13 ...	69896c7d-e68e-...	0,033	e3d883ca-1e19-4217-9dcb-c79615e5e296
2007.01.13 ...	69896c7d-e68e-...	0,001	f863729d-84a2-42d2-8930-e033260177b1
2007.01.13 ...	69896c7d-e68e-...	0,0175	a0407c84-2e0d-43a5-bb82-e0456e01302e
2007.01.13 ...	69896c7d-e68e-...	0,0191000011	55a8a163-bf5c-4118-9fe8-e0973df33bb6
2007.01.13 ...	69896c7d-e68e-...	0,022499999	983c5b44-d0d5-4b97-9f47-e249d1c900b6

9.4 Priedas Nr.4 *Straipsnis konferencijos „Informacinė visuomenė ir universitetinės studijos 2007“ leidinyje*

PORTALO STATISTIKOS MODULIS PAGRĮSTAS GRUPAVIMU

Martynas Ruzgys

Kauno technologijos universitetas, Informacijos sistemų katedra, Studentų g. 50, Kaunas

Pristatomas duomenų gavybos ir grupavimo naudojimas paplitusiose sistemose bei sukurtas IT žinių portalo statistikos prototipas duomenų saugojimui, analizei ir peržiūrai atlikti. Siūlomas statistikos modulis duomenų saugykloje periodiškais laiko momentais vykdantis duomenų transformacijas. Portale prieinami statistiniai duomenys gali būti grupuoti. Sugrupuotą informaciją pateikus grafiškai, duomenys gali būti interpretuojami ir stebimi veiklos mastai. Panašių objektų grupėms išskirti pritaikytas vienas iš žinomiausių duomenų grupavimo metodų – lygiagretusis k-vidurkių metodas.

ĮVADAS

IT žinių portalo paskirtis – kaupti mokslinius IT srities straipsnius dalyvaujant vartotojams-autoriams, tuo pačiu straipsnius bei autorius reitinguoti, vertinti, taip skatinant portalo plėtrą. Pagal sukauptą portalo statistiką vartotojas gali stebėti portalo veiklos mastus, o administratorius pagal tai gali priimti sprendimus dėl sisteminių apribojimų (reitingų lygių ribos, maksimalios reikšmės reitingų sudėties elementams ir kt.) keitimo. Portalo statistikos naudotojas pirmiausiai nori gauti kuo daugiau informatyvumo, todėl svarbu yra informacijos pateikimo lankstumas: procentiniai palyginimai, pasirinkimo galimybės, įvairūs kiekiai, kritiniai rodikliai ir pan.

Sukurtas IT žinių portalo statistikos prototipas duomenų saugojimui, analizei ir peržiūrai atlikti gali būti pavadintas duomenų gavybos (DG) sistema. Ji padeda atrinkti duomenis iš didelio jų kiekio ir suprantamai pateikti. Duomenims atrinkti ir panašių objektų grupėms išskirti pritaikytas vienas iš žinomiausių duomenų grupavimo metodų – lygiagretusis k-vidurkių metodas. Statistikos modulis leidžia peržiūrinėti objektų (straipsnių, autorių) veiksmų ar veiksmų su objektais statistiką laike.

Portalas suprojektuotas su specifine duomenų baze (DB), todėl sudėtinga pritaikyti jau sukurtą, universalią statistikos sistemą siauros srities duomenims kaupti ir analizuoti. Informacija daugumoje sprendimų yra saugoma periodiškais laiko momentais, o realaus laiko sistemos reikalauja didelio sistemos pajėgumo, nes būna labai apkrautos. Todėl sistemoje periodiškais laiko momentais vykdomos transformacijos, kai portalo sukaupti duomenys apibendrinami, diskretizuojami ir saugomi statistikos reikmėms.

DG įrankiai perauga į prognozuojančiosios analizės sistemas, kurios paskutiniu metu prilipdė rinką [22]. Parinkus prognozavimo modelio sudarymo metodiką, galima gauti sistemą, tokią kaip STATISTICA Data Miner ar StarProbe Data Miner, kuri prognozuotų tam tikrą įvykio eigą pagal suklasifikuotus duomenis.

STATISTIKOS MODELIŲ IR METODŲ ANALIZĖ

Duomenų gavyba ir grupavimas

„DG yra prasmingų dėsningumų, modelių ir tendencijų radimo procesas dideliuose informacijos kiekiuose, naudojant modelių atpažinimo statistinius bei matematinius metodus“ [1]. DG yra procesas skirtas automatiškai aptikti šablonus didelės apimties duomenyse naudojant tokias priemones, kaip asociacijų paieška, eiliškumų analizė, grupavimas, klasifikacija, įvertinimas, prognozavimas. Kiekvienas iš šių metodų sprendžia tam tikras problemas [2].

Duomenų grupavimas – tai objektų klasifikavimas į skirtingas grupes, tiksliau tariant, duomenų dalinimas į pogrupius, kad kiekviename jų duomenys turėtų bendrų bruožų – dažniausiai tai panašumas pagal numatytą atstumo matą. Duomenų grupavimas yra dažnas metodas statistiniam duomenų tyrimui. Metodas naudojamas tokiose srityse, kaip save mokančios sistemos, duomenų gavyba, šablonų atpažinimas, vaizdų analizė, bioinformatika ir genų inžinerija, statistika [3].

Šiai sričiai aktualūs grupavimo metodų tipai yra du: hierarchinis ir padalijimo.

Hierarchiniai metodai grupę formuoja hierarchiškai, t.y. kiekviena grupės viršūnė turi vaikinę grupę. Grupės apjungiamos ir skaidomos, taip sudarant hierarchinę struktūrą. Priskiriami klasikiniai SLINK, COBWEB algoritmai bei

naujesni CURE ir CHAMELEON algoritmai. Išskiriami sujungimo (Agglomerative) ir išskaidymo (Divisive) algoritmų tipai.

Padalijimo (Partitioning) metoduose duomenys dalinami į kelis pogrupius. Visų pogrupių patikrinti neįmanoma, todėl naudojama iteracinė optimizacija, kuri palaipsniui gerina grupes.

Išskiriami tipai:

- Kelties (Relocation) metodai;
- Tikimybinio grupavimo metodai: EM šablono, SNOB, AUTOCLASS, MCLUST;
- Artimiausių kaimynų grupavimo metodai;
- K-vidurinių taškų (K-medoids) grupavimo metodai: PAM, CLARA, CLARANS;
- K-vidurkių (K-means) grupavimo metodai;
- Tankumo (Density-Based) metodai: DBSCAN, OPTICS, DBCLASD, DENCLUE; [4]

DG technologijos paremtos objektų panašumo matais. Atstumo vadinamas duomenų nepanašumas, o atstumas yra matuojamas. Žinomiausias atstumo matas yra Euklido atstumas:

$d_E(x_i, x_j) = \sqrt{\sum_{k=1}^{m} (x_{ik} - x_{jk})^2}$, Kur x_i, $x_j \in X$, $(i,j = 1, \ldots, n)$, $x_i = \{x_{i1}, x_{i2}, \ldots, x_{im}\}$ – duomenų vektorius, m – dimensijų skaičius, n – vektorių skaičius;

Euklido kvadratinių paklaidų principu kiekviena grupė atstovaujama prototipinio objekto, o kiti priskiriami grupėms pagal panašumą (kuo mažesnis atstumas) [5].

K-vidurkių grupavimo metodas

K-vidurkių algoritmas - tai vienas iš paprasčiausių, žinomiausių ir dažniausiai naudojamų algoritmų duomenų grupavimui. Jis paremtas kvadratinių paklaidų kriterijumi. Šis algoritmas pateikia efektyvius grupavimo rezultatus daugelyje praktinių taikymo sričių: kalbos bei vaizdų atpažinimui, genetikos, bioinformatikos, geografinių duomenų analizei ir kt. [6].

Trūkumai:

- Sudėtingumas stipriai veikiamas duomenų objektų kiekio, nes reikia perskaičiuoti atstumus tikrinant visus objektus pakartotinai ir iteratyviai juos pergrupuoti. Kai duomenų kiekis labai didelis, algoritmas užtrunka ilgiau laiko skaičiavimams, o tai didina skaičiavimo sudėtingumą ir mažina skaičiavimo greitį duomenų rinkiniuose.
- Algoritmo jautrus pradiniam grupių kiekiui. Jei šis parenkamas netinkamai, kriterijaus funkcija gali konverguoti į lokalų minimumą. Algoritmui reikėtų mažiau iteracijų, jei pradiniai centrai būtų parenkami pagal geresnę strategiją [9].

Lygiagretusis k-vidurkių algoritmas ypač pagerina neefektyviausią paprastojo k-vidurkių algoritmo iteratyvaus pergrupavimo dalį. Pagal jį duomenys sudalinami į m lygiagrečių procesų, kurie skaičiuoja atstumus tarp kiekvieno duomenų objekto ir k grupių centrų. Objektas priskiriamas artimiausiai grupei. Kiekvienai grupei surenkama informacija iš lygiagrečių procesų ir perskaičiuojami nauji grupių centrai [8].

Parastojo k-vidurkių algoritmo skaičiavimo sudėtingumas – $O(ktn)$, kur n – duomenų objektų skaičius, k – grupių skaičius, t – iteracijų skaičius. Lygiagrečiojo k-vidurkių algoritmo sudėtingumas sumažinamas m kartų [7].

Duomenų grupavimo taikymas statistikos sistemoms

Mineset (SGI) – DG ir analizės sistema

MineSet įrankių rinkinys leidžia išgauti, analizuoti ir grafiškai atvaizduoti duomenis, kad juos būtų galima suprasti ir tirti.

Gavyba

Naudojami analitiniai DG algoritmai, kurie sukuria modelį iš duomenų. DG įrankiai automatiškai suranda šablonus, atranda tendenciją ir sukuria modelį. Dažniausiai naudojamos dvi modeliavimo algoritmų šeimos: prižiūrimasis ir neprižiūrimasis modeliavimas.

Prižiūrimosios užduotys yra prognozuojamojo modeliavimo užduotys, kurių tikslas yra numatyti vieno stulpelio reikšmę pagal kitų stulpelių reikšmes. Dažniausios šio modelio užduotys yra klasifikacija ir regresija.

Neprižiūrimųjų užduočių tikslas – atrasti šablonus ir duomenų dalis pagal elgsenos panašumą, tai – aprašymo, o ne prognozavimo užduotis. MineSet teikia du šio tipo modeliavimo metodus: asociacijų (sąsajos taisyklių nustatymas tarp kelių duomenų atributų) ir grupavimo (duomenis skaidymas į įrašų grupes pagal panašias charakteristikas).

Grupavimas

Naudojami du grupavimo algoritmai: k-vidurkių, iteratyvusis k-vidurkių. Įrankis turi daug patogių pasirinkimų: grupavimo modelio, minimalaus-maksimalaus iteracijų skaičiaus, grupių kiekio ar intervalo, atstumo mato, atributo svorio, atsitiktinių pradinių reikšmių generavimo [10].

STATISTICA Data Miner (StatSoft) – DG, analizės, prognozavimo sistema

Įrankio paskirtis – atskleisti paslėptas tendencijas, paaiškinti žinomas struktūras, prognozuoti elgseną.

STATISTICA Data Miner apima daug pilnai integruotų pažangių DG ir prognozavimo metodų.

Gavyba

Platus naudojamų gavybos metodų pasirinkimas: daugiklių analizė (Factor Analysis), bendri klasifikavimo ir regresijos medžių modeliai (GTrees), naudojantys CART metodo realizaciją, bendrieji tiesiniai ir regresijos modeliai (GLM, GLZ, GRM), bendrasis diskriminanto analizės modelis (GDA), dalinių kvadratinių paklaidų (PLS), CHAID modeliai, naudingi tiriamajai duomenų analizei ir prognozavimui.

Grupavimas

Iš grupavimo metodų šis įrankis naudoja medžių grupavimo, k-vidurkių grupavimo analizę, diskriminanto analizės modelius ir praplėstą tikimybės maksimizavimo EM metodą su v-fold maišymo ir kryžminio tikrinimo pasirinkimu optimizuotam grupių kiekiui. Šie grupavimo metodai skirti apdoroti dideliems duomenų rinkiniams, juos grupuojant ir suteikiant pagrindą šablonų atpažinimui. Pažangūs EM grupavimo metodai siejami su tikimybinio, statistinio grupavimo galimybėmis [11].

StarProbe Data Miner (Rosella) – duomenų grupavimo, analizės, prognozavimo sistema

Įrankio paskirtis – neuroninis grupavimas, duomenų analizė ir tikimybinis modeliavimas.

Grupavimas

StarProbe palaiko platų grupavimo algoritmų pasirinkimą: neuroninis grupavimas – naujoviškas įrankis grupavimui, naudojantis neuroninius tinklus, save organizuojančius žemėlapius (SOM); tikimybinis segmentacijos modeliavimas – tinklas išmokomas grupavimo šablonų ir tada taikomas duomenų segmentacijai ir statistinių reikšmių prognozavimui; karštų taškų analizė; kryžminė lentelių analizė ir kt. [12].

Mineset, STATISTICA ir StarProbe įrankiai naudoja tradicines ir pažangias atvaizdavimo priemones (medžių atvaizdavimas, išsibarstymo, 3D diagramos ir pan.).

Internetiniai statistikos pateikimo sprendimai

Buvo apžvelgta kelių lietuviškų portalų statistikos dalis ir jos pateikimo būdai.

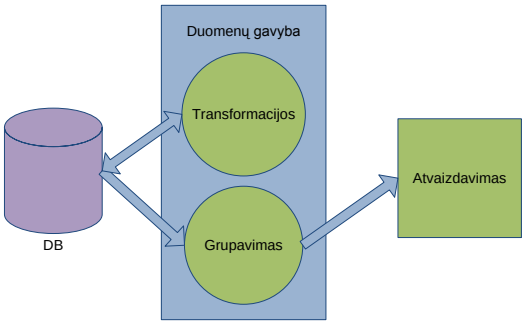
Dažniausiai pateikiami paprasti duomenų sąrašai su kiekiais, vidurkiais ir procentinėmis reikšmėmis, kur to reikia, o rezultatai atvaizduojami ne grafikais, o tiesiog gražiai apipavidalintomis lentelėmis [13][20].

Įdomesnis statistikos pateikimas [21], kur naudojama modifikuota histograma pageidaujamam atlyginimui skaičiuoti. Pasirinkus filtro reikšmes pateikiami rezultatai. Nežinomu metodu parenkami pradinis ir galinis rėžiai bei žingsnis. Pastebima, kad histogramos stulpelių kiekis priklausomas nuo nagrinėjamų žmonių skaičiaus. Kiekviename stulpelyje matomas patekimo į jį dažnis (žmonių kiekis ir procentai).

STATISTIKOS MODULIO KONCEPTUALIOJI SPECIFIKACIJA

Statistikos modulio koncepciją galima pavaizduoti (**1 pav.**) kaip DB pirmoje grandyje, DG priemones, kontaktuojančias su DB, vidurinėje grandyje ir atvaizdavimo priemones paskutinėje grandyje. Duomenų transformacijos vyksta kasdien, o to proceso metu portalo lentelių duomenys diskretizuojami, apibendrinami ir atitinkamais pjūviais saugomi statistikai skirtose lentelėse. Grupavimas („KMeans“ komponentas) ir atvaizdavimas

(„DotNetCharting“ komponentas) vyksta vartotojui naršant ir pasirenkant peržiūras, reikalaujančias grupavimo. Rezultate gaunami atvaizduoti duomenys.



4 pav. **Statistikos modulio koncepcinė schema**

Funkcionalumo pjūviai

Vartotojams pateikiamas funkcionalumas dviem požiūriais: straipsnių ir autorių. Kiekvienam šiam pjūviui galimos statistikos peržiūros:

- bendra (straipsnių/autorių rašymo, skaitymo informacija);
- vertinimų (straipsnių vertinimų, autoriaus parašytų straipsnių vertinimų, autoriaus atliktų vertinimų);
- recenzavimo;
- reitingų;

O šiose peržiūrose priklausomai nuo vartotojo tipo (lankytojas ar registruotas autorius) matoma atitinkamai apibendrinta arba detali informacija. Grupuota informacija, leidžianti matyti veiklos mastus, matoma tik registruotiems vartotojams.

Duomenys ir jų transformacijos

Statistikos modulis naudojasi viso portalo DB ir saugo diskretizuotus duomenis kasdien. Įrašai apie vykstančius įvykius fiksuojami portalo veiklos lentelėse, o statistinei informacijai kaupti iš jos paimami ir keliais pjūviais saugomi statistikos lentelėse. Statistikos modulio DB modelyje skirtos atskiros lentelės straipsniams ir autoriams pagal kiekvieną veiksmą (straipsnių vertinimų, autorių vertinimų, straipsnių recenzijų, autorių recenzijų, straipsnių skaitymų, straipsnių rašymo, reitingų). Visa smulki informacija nėra reikalinga statistiniam analizavimui, todėl duomenų diskretizuojami [14]. Tai vyksta serveryje saugomų procedūrų pagalba, kurios apdoroja duomenis iš bendrų portalo lentelių ir saugo juos į statistikos lenteles. Duomenys apdorojami dažniausiai dviem aspektais: straipsnio ir autoriaus. Serverio valdymo įrankiuose sukuriama užduotis (job), kuri kasdien kreipiasi į portalo veiklos duomenis diskretizuojančias procedūras.

Diena laikoma mažiausiu nedalomu laiko vienetu statistikos duomenų peržiūroje. Nagrinėti įvykius valandomis ar smulkiau visiškai nereikia, nes pati portalo veikla ir jo duomenų analizė nėra kritinio svarbumo kaip medicininių, akcijų rinkos ar pan. duomenų analizė.

Vertinimo duomenų apdorojimo procedūra apibendrina dienos vertinimų duomenis ir informaciją autoriaus bei straipsnio požiūriu saugo į atitinkamą statistikos lentelę. Kaupiama tik vertinimo statistika per nurodytą dieną. Taip pat saugomi duomenys apie straipsnio sukauptas vertinimo reikšmes ir autoriaus vertinimus, sukauptus kasdien nuo pat autoriaus/straipsnio atsiradimo portale. Jei kažkurią dieną yra autoriaus vertinimas, tai į tą dieną atkeliami seniau sukaupti vertinimai.

Recenzavimo duomenų apdorojimo procedūra apibendrina dienos recenzijų duomenis, t.y. autoriaus požiūriu saugo duomenis apie jo recenzijų kiekius per dieną, o straipsnio požiūriu – apie straipsniui parašytas recenzijas per dieną.

Reitingavimo duomenų apdorojimo procedūra, pasikeitus einamojo periodo svoriui (tai būna labai retai), saugo periodo svorio reikšmes. Taip pat kasdien saugo einamajame periode sukauptas autorių/straipsnių reitingų reikšmes ir sukauptas autorių/straipsnių reitingų sudėtinių elementų reikšmes.

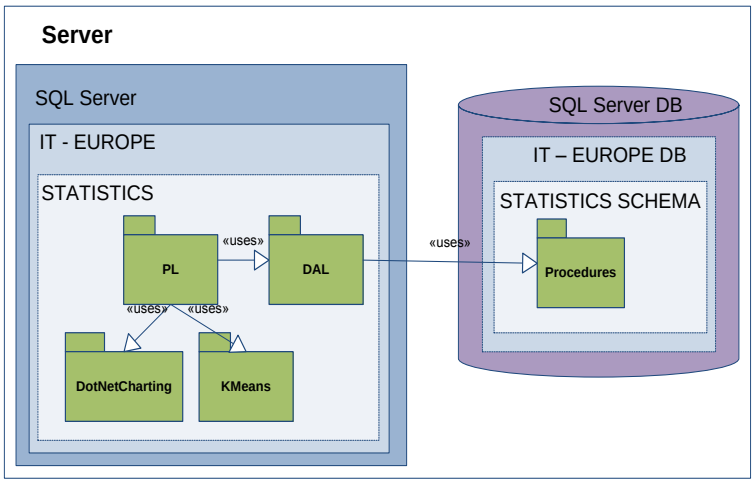
Straipsnių skaitymo duomenų apdorojimo procedūra apibendrina dienos skaitymų duomenis ir straipsnių požiūriu saugo jo skaitymų kiekius per dieną. O straipsnių rašymo duomenų apdorojimo procedūra apibendrina dienos rašymų duomenis ir autorių požiūriu sumuoja kiekviena kalba straipsnių kiekius, prie kurių kūrimo jis prisidėjo.

Statistikos modulio architektūra

Aukščiausiame lygyje sistemą sudaro penkios dalys (komponentai):

- **„KMeans“** – Grupavimo komponentas;
- **„DotNetCharting“** – Komponentas grafiniam atvaizdavimui;
- **„PL“** – Atvaizdavimo ir valdymo dalis;
- **„DAL“** – Prieigos prie duomenų lygis;
- **„Procedures“** – Serveryje saugomos procedūros darbui su duomenimis;

Išnaudotas MS.NET Framework 2 siūlomas privalumas atskirti atvaizdavimo lygį (PL), duomenų prieigos lygį (DAL) bei serveryje saugomas procedūras (**2 pav.**).



5 pav. **Statistikos modulio architektūra**

Komponento „KMeans“ detalizavimas

Greitesniam programinė įrangos vystymui pasiekti pasinaudota pakartotiniu komponentų panaudojimu [16]. Pritaikytas jau realizuotas (C# programavimo kalba) lygiagretusis k-vidurkių algoritmas [17].

Naujai bandyta pažvelgti į anksčiau įvardintą paprastojo k-vidurkių algoritmo trūkumą dėl pradinių centrų parinkimo geresnės strategijos, kad algoritmui reikėtų mažiau iteracijų. Todėl algoritmas buvo papildytas. Strategija – pradiniams grupių centrams iš eilės imti duomenų rinkinio nesikartojančius duomenis.

Komponentą sudaro klasės, realizuojančios kelių matų duomenų grupavimą k-vidurkių algoritmu, pagal Euklido atstumą. „Kmeans“ komponento duomenų rinkinio grupavimo pagrindinio metodo veikimas:

1. Į grupių kolekciją įdedamas pradinis nesikartojančių duomenų kiekis lygus pasirinktam grupių kiekiui.
2. Vyksta grupių kolekcijos perskaičiavimas. Tam duomenys išskirstomi į *m* lygiagrečių procesų, kuriuose randama duomenų dalies vieta (pagal Euklido atstumą) naujoje grupių kolekcijoje. Duomenys iš lygiagrečių procesų surenkami ir susumuojami į rastas artimiausias grupes.
3. Skaičiuojamas stabilių grupių kiekis, t.y. naujos ir senos grupių kolekcijos atitinkamų grupių, tarp kurių atstumas lygus nuliui, kiekis.
4. Jei visos grupės stabilios, gaunamas rezultatas, priešingai – kartojamas 2 žingsnis.

Čia funkcionuoja objektas „Cluster“, t.y. grupė, kuri turi visų į ją patenkančių duomenų vidurkį ir sumą. Klasė „KMeansParallel“ sukuria grupių kolekciją, jai priskiriamos pradinės reikšmės. Duomenų rinkinys išdalinamas lygiagretiems procesams. Procesai kreipiasi į delegatą „ClusterPartialDataSetDelegate“, kuris pagal Euklido atstumą nustato, kuriai grupei duomenys artimesni. „KMeansParallel“ surenka duomenis iš lygiagrečių procesų ir pagal tai perskaičiuoja grupių kolekcijos centrus.

Komponento „DotNetCharting“ detalizavimas

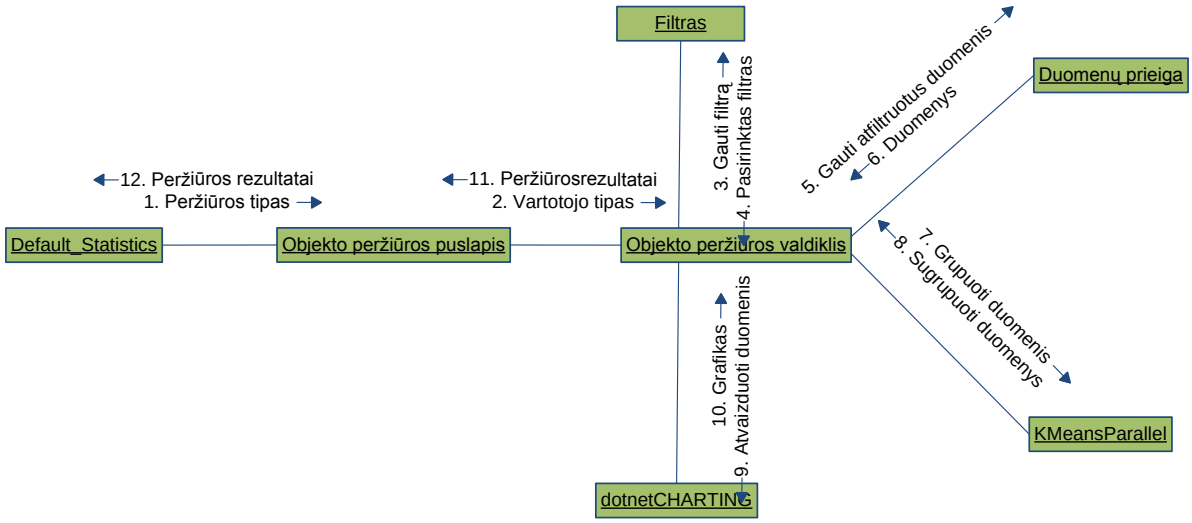
Panaudotas grafinio atvaizdavimo komponentas „DotNetCharting“ (WebAvail Porductions, Inc. & Corporate Web Solutions Ltd.) [15]. Realizuotas .NET(C#) priemonėmis. Jį patogu naudoti ASP.NET(C#/VB) ar Windows formų programose.

Turi duomenų paėmimo galimybes tiesiogiai iš duomenų bazės (MySQL, ACCESS, SQL Server, Oracle, ODBC), Excel ar XML failų bei programinių objektų (DataSet, DataTable, DatView, ArrayList, Object Collection, HashTable). Komponentas naudojamas kaip .DLL biblioteka. Kreipiantis į komponentą, jam perduodami duomenys, o gaunamas duomenų atvaizdavimas diagrama.

Komponento „PL“ detalizavimas

Pakete pateikiamos klasės, realizuojančios registruoto ir neregistruoto lankytojo sąsają, bendrą puslapių struktūrą, kitų komponentų valdymą.

3 pav. pavaizduota bendradarbiavimo diagramos principas, tinkantis visiems panaudojimo atvejams, kur naudojamas grupavimas ir atvaizdavimas. Pagrindinis puslapis „Default_statistics“ kviečia vieną iš straipsnių arba autorių peržiūros komponentų (iš viso 7), o šis pagal vartotojo tipą kreipiasi į žemesnį peržiūros tipo valdiklį. Pastarieji apdoroja filtrų sukūrimą ir pasirinkimą, kreipimąsi į serverio procedūras per „DAL“ informacijai gauti. Valdikliai pagal poreikį sukuria reikiamą kiekį „KMeansParallel“ bei „DotNetCharting“ komponentų, kreipiasi į juos atitinkamai informacijai grupuoti arba grupuotai informacijai atvaizduoti.



6 pav. **Apibendrinta bendradarbiavimo diagrama duomenų grupavimui ir atvaizdavimui**

Komponento „DAL“ detalizavimas

„DAL“ – Duomenų prieigos lygis. Čia sukauptas visas specifinis kodas, skirtas prieiti duomenims iš DB bei prisijungimo prie DB nustatymai. Naudojamas MS.NET Framework *DataSet* objektas, *TableAdapters* (klasė su metodais darbui tarp *DataSet* objekte esančių *DataTable* ir duomenų bazės) [18].

Atvaizdavimo lygis „PL“ turi kreiptis į „DAL“ metodus darbui su norimos lentelės duomenimis.

Komponento „Procedures” detalizavimas

„Procedures“ – tai MSSQL serverio duomenų bazėje saugomos T-SQL procedūros. Saugomų procedūrų nauda:

- Sukompiliavimas iš anksto atsiperka, kai procedūros naudojamos pakartotinai;
- Sumažintas apkrovimas serveryje;
- Gali būti naudojamos kelių programų ar vartotojų;
- Padidinta saugumo kontrolė leidžia suteikti teises vartotojams vykdyti procedūras nepriklausomai nuo lentelės teisių; [19]

„DAL“ turi prieigą prie šių procedūrų, kurios naudojamos komponente „PL“ darbui su duomenimis Iš viso ~38 serveryje saugomos procedūros.

IŠVADOS

Sukurta sistema skirta siauros srities portalui, kur galingų ir daug funkcijų teikiančių grupavimo įrankių nereikia. Pakanka pasirinko vieno efektyvaus grupavimo algoritmo. Tam naudotas jau realizuotas lygiagretusis k-vidurkių grupavimo algoritmas [17], kuris buvo pakoreguotas pagal poreikį ir pritaikytas sistemoje. Todėl grupavimas veikia tiksliai.

Statistikos modulį galima pavadinti DG sistema, nes ji leidžia iš portalo DB sukauptų duomenų išgauti apibendrintą informaciją, kurios pobūdį visada galima praplėsti tobulinant sistemą. Išgautą informaciją pateikus grafiškai, galima interpretuoti duomenis ir stebėti veiklos mastus ar tendencijas. Vienas iš rinkoje siūlomų grafinio atvaizdavimo įrankių - „DotNetCharting“ [15]. Jį galima naudoti nemokamai.

Statistikos modulio realizacijai naudota MSSQL serverio duomenų bazės valdymo sistemą (DBVS), kurią diktuoja visas portalas. Realizacijai taip pat parinkta Microsoft Visual Studio .Net 2005 programavimo aplinka, nes geriausiai dera prie naudojamos DBVS. ASP.NET(C#) technologija puikiai tinka internetinėms aplikacijoms kurti, o prieigai prie duomenų bazių patogi ADO.NET technologija. Panaudotas MS.NET Framework 2 siūlomas duomenų prieigos lygio atskyrimas.

Naudotas komponentų derinys pigus kūrimo kaštų atžvilgiu, kadangi remiamasi patogia pakartotinio panaudojimo technologija. Todėl visų komponentų kurti iš naujo tikrai neapsimoka.

LITERATŪROS SĄRAŠAS

[1] **Gartner Group**. Data Mining, [žiūrėta 2006-11-30]. Prieiga per internetą: http://www2.nr.no/documents/samba/research_areas/BAMG/Pattern/datamining.html

[2] **Wikipedia**. Data Mining, [žiūrėta 2007-02-29]. Prieiga per internetą: http://en.wikipedia.org/wiki/Data_mining

[3] **Wikipedia**. Data Clustering, [žiūrėta 2007-03-01]. Prieiga per internetą: http://en.wikipedia.org/wiki/Data_clustering

[4] **P. Berkhin**. Survey of Clustering Data Mining Techniques, *Accrue Software*, 2002

[5] **M. Kantardzic**. Data Mining: Concepts, Models, Methods and Algorithms, *ff*, 2003

[6] **Wikipedia**. K-means algorithm, [žiūrėta 2007-03-05]. Prieiga per internetą: http://en.wikipedia.org/wiki/K-means_algorithm#Demonstration_of_the_algorithm

[7] **T. Kanungo, D. M. Mount, N. S. Netanyahu, Ch. D. Piatko, R. Silverman, A. Y. Wu**. An Efficient k-Means Clustering Algorithm: Analysis and Implementation, *from IEEE transactions on pattern analysis and machine intelligence*, vol. 24, No. 7, Liepa 2002

[8] **T. Jinlan, Z. Lin, Z. Suqin, L. Lu**. Improvement and Parallelism of *k*-Means Clustering Algorithm, *Tsinghua science and technology*, ISSN 1007-0214 01/21 277-281 pslp, Volume 10, Number 3, Birželis 2005

[9] **S. Kantabutra, C. Naramittakapong, P. Kornpitak**. Pipelined K-means Algorithm on COWs, *The Theory of Computation Group*, 2003

[10] **SGI**, Mineset, [žiūrėta 2007-03-10]. Prieiga per internetą: http://techpubs.sgi.com/library/tpl/cgi-bin/getdoc.cgi/0650/bks/SGI_EndUser/books/MineSetNT_T/sgi_html/pr01.html; http://www.sgi.com/

[11] **StatSoft**, STATISTICA, [žiūrėta 2007-03-18]. Prieiga per internetą: http://www.statsoft.com/products/products.htm

[12] **Rosella**, StarProbe, [žiūrėta 2007-03-10]. Prieiga per internetą: http://www.roselladb.com/starprobe.htm

[13] **System-admins**, Portalas IT sistemų administratoriams. [Žiūrėta 2005-11-25]. Prieiga per internetą: http://www.system-admins.net

[14] **J. Gama, C. Pinto**. Discretization from Data Streams: Applications to Histograms and Data Mining, *LIACC, FEP, University of Porto*, 2004

[15] **WebAvail Porductions, Inc. & Corporate Web Solutions Ltd**, Grafinio atvaizdavimo komponentas DotNetCdarting, [žiūrėta 2007-02-20]. Prieiga per internetą: http://www.dotnetcharting.com

[16] **I. Sommerville**, Software Engineering, 6th edition. Chapter 14, 2000

[17] **Professional Community Server**, Free Data Mining Source Code, [žiūrėta 2007-04-30]. Prieiga per internetą: http://www.kdkeys.net/forums/6051/ShowThread.aspx

[18] **Microsoft Corporation**, Creating a Data Access Layer, [žiūrėta 2006-01-15]. Prieiga per internetą: http://www.asp.net/learn/dataaccess/tutorial01cs.aspx?tabid=63

[19] **M. Chapple**, Stored Procedures in SQL Server, [žiūrėta 2006-12-05]. Prieiga per internetą: http://databases.about.com/od/sqlserver/l/aastoredprocs.htm
[20] **INTERPREKYBA**, Klasiokų bendravimo portalas. [Žiūrėta 2006-11-25]. Prieiga per internetą: http://www.klase.lt
[21] **MediaWorks**, Darbo paieškos sistema. [Žiūrėta 2006-11-20]. Prieiga per internetą: http://www.cv.lt
[22] **L. Agosta**. The Future of Data Mining – Predictive Analytics, *DM Review Magazine*, Rugpjūtis 2004

PORTAL STATISTICS MODULE BASED ON CLUSTERING

Presented data mining methods and clustering usage in current statistical systems and created statistics module prototype for data storage, analysis and visualization for IT knowledge portal. In suggested statistics prototype database periodical data transformations are performed. Statistical data accessed in portal can be clustered. Clustered information represented graphically may serve for interpreting information when trends may be noticed. One of the best known data clustering methods – parallel k-means method – is adapted for separating similar data clusters.