KAUNO TECHNOLOGIJOS UNIVERSITETAS
INFORMATIKOS FAKULTETAS
KOMPIUTERIŲ KATEDRA

Eduardas Andrėkus

Portalo duomenų apdorojimo ir analizės sistema

Magistro darbas

Darbo vadovas
prof. E. Kazanavičius

Kaunas, 2006

KAUNO TECHNOLOGIJOS UNIVERSITETAS
INFORMATIKOS FAKULTETAS
KOMPIUTERIŲ KATEDRA

Eduardas Andrėkus

Portalo duomenų apdorojimo ir analizės sistema

Informatikos mokslo magistro baigiamasis darbas

Kalbos konsultantė
Lietuvių k. katedros lekt.
I.Mickienė
2006-05-29

Vadovas
prof. E. Kazanavičius
2006-05-29

Recenzentas
doc. dr. E. Toldinas
2006-05-29

Atliko
IFM 0/1 gr. stud.
E. Andrėkus
2006-05-29

Kaunas, 2006

Portal data processing and analysis system

SUMMARY

Web analysis is important task in web portals development. Web measurement can answer to many questions related to user activity. Web analytics is the study of whether the web site is meeting its diverse goals. Just as important, it is the presentation of the results to the various divisions of the company in a comprehensible format. The core of the web analytical system is built on traditional decision support technologies that are focused specifically on capturing, analyzing and reporting on web visitor data.

In this work analyzed methods ant technologies required for web portal analysis system design. This paper compares possible data gathering sources and methods for reports creation. There is offered conceptual design of analysis system using hybrid data collection sources. System realization is used in KTU Computer Cathedral activity.

TURINYS

1. Įvadas

Pastaruoju metu sparčiai skverbiantis internetui į kasdienį daugelio žmonių gyvenimą, vis intensyviau vystosi internetu teikiamos paslaugos. Vis daugiau ir daugiau laiko praleidžia žmonės ieškodami internete informacijos, prekių ar paslaugų. Tuo tarpu interneto svetainių savininkai nori pritraukti kuo daugiau lankytojų. Taigi iškyla poreikis kuo geriau suprasti lankytojų elgseną, surasti kuo geresnes pritraukimo į svetainę priemones, sužinoti ar lankytojai yra patenkinti svetainės galimybėmis. Svarbu ne tik, informacijos pateikimas, bet ir tai, kad ji turi būti lengvai pasiekiama. Apie tai galima sužinoti tik analizuojant lankytojų elgseną. Pradžioje, analizuojant tinklapių lankomumą, būdavo fiksuojamas tik bendras tinklapyje apsilankiusių vartotojų skaičius. Tyrinėjant detaliau tokios informacijos nebeužteko. Yra svarbus bendras apsilankymų skaičius. Tačiau paprasta statistika, kiek buvo paprasčiausių apsilankymų, suteikia labai mažai informacijos, todėl reikalinga daug gilesnė analizė. Svarbu analizuoti, kaip atėjęs į portalą elgiasi vartotojas, kiek laiko jis užtrunka viename puslapyje, kokiomis nuorodomis dažniausiai naudojasi, kokiu eiliškumu dažniausiai peržiūri informaciją.

Nors portalo duomenų apdorojimo ir analizės sistemos nėra naujas dalykas informatikos srityje, vis dėlto yra nemažai vietos tobulėjimui, žinant kad šios srities rinka sparčiai plečiasi. Projektuojant tokias sistemas, reikia įvertinti daugelį kriterijų – tiek bendrų, tiek atsižvelgiant į unikalias portalo savybes. Yra sukurtos įvairios tinklapių ir portalų duomenų apdorojimo ir analizės priemonių. Tačiau šios priemonės dažniausiai yra pritaikytos konkrečiam atvejui. Nemažai jų būna jau sukurto portalo dalis, todėl jų panaudoti savom reikmėm yra praktiškai neįmanoma. Kuriant savo portalą šios priemonės yra netinkamos ir reikia kurti savas. Yra nemažai analizės sistemų, kurios kaip paslaugos parduodamos internete. Bandant jas prisitaikyti prie sukurto portalo kyla problema, nes reikia atlikti esamų produktų analizę ir išsiaiškinti, kuri iš siūlomų sistemų labiausiai atitinka poreikius. Taigi atsiranda poreikis kurti metodiką, kaip ,kuriant portalą, reiktų surinkti ir analizuoti jo lankytojų elgsenos duomenis, į kokius kriterijus reikia atsižvelgti, kuriant savo ar renkantis portalo duomenų apdorojimo ir analizės sistemą.

Tyrimo sritis apima KTU kompiuterių katedros studijų modulių portalo lankomumo duomenų apdorojimą ir analizę.

Darbo tikslas – išnagrinėti duomenų rinkimo ir analizės metodus portaluose ir jų pagrindu sukurti universalią portalo duomenų apdorojimo ir analizės metodiką. Išnagrinėti priklausomybę, tarp sistemos kompiuterinių resursų apkrautumo ir išanalizuotos informacijos ataskaitų pateikimo greičio. Apžvelgti, kokie galimi duomenų analizavimo būdai, bei kokios

galimos ataskaitos. Darbo tikslas yra sukurti sistemą, leidžiančią stebėti, ar portale pateikiama informacija yra skaitoma, taip pat analizuoti kuriose portalo puslapiuose yra didžiausias apkrautumas.

Uždaviniai. Pagrindinis darbo uždavinys yra sukurti portalo duomenų apdorojimo ir analizės metodiką. Tam kad pasiketi užsibrėžtą tikslą reikia išspręsti šiuos tarpinius uždavinius:

- Išanalizuoti literatūroje minimus galimus duomenų rinkimo šaltinius, bei galimus analizavimo metodus.
- Apžvelgti kriterijus, kuriais remianti yra analizuojami portalai.
- Sukurti universalią portalo duomenų apdorojimo ir analizės metodiką.
- Išanalizuoti KTU kompiuterių katedros studijų modulių portalą: nustatyti galimus duomenų rinkimo šaltinius ir poreikį ataskaitoms.
- Realizuoti, remiantis sukurta metodika, duomenų apdorojimo ir analizės sistemą KTU kompiuterių katedros studijų modulių portale.

2. Portalo duomenų rinkimo ir analizės metodų analizė

2.1 Portalo duomenų apdorojimo ir analizės sistemos samprata

Portalas – dažniausiai apibrėžiamas kaip svetainė, pasižyminti didele informacijos bei papildomų paslaugų gausa, kuri gali tarnauti kaip išeities taškas į kitus interneto resursus. Portalai paprastai pateikia kombinuotą, personalizuotą prieigą prie informacijos, duomenų ir funkcionalumų bei pasižymi dideliu informacijos ir funkcionalumų integravimo lygmeniu. Tinklapių analizė (*Web analysis*) yra įvairių duomenų įvertinimas, įskaitant: tinklapių srautą, tinklapiuose vykstančių transakcijų, tinklapių serverio našumą, naudojimo studijos, vartotojų persiųstos informacijos ir kitų susijusių šaltinių pagalba sukurti apibendrintą supratimą apie vartotojų patirtį tinklapyje [3]. Portalo duomenų apdorojimo ir analizės sistemos turi pasižymėti didele duomenų integracija tarp portalo ir analizės duomenų. Turi būti galimybė daug didesnė analizuojamos informacijos detalizavimas (pvz. Leisti analizuoti konkretaus vartotojo veiksmus).

Portalo analizė yra svarbi, nes jos metu galima sužinoti ar portalo naudojimas pateisina lūkesčius. Dauguma organizacijų analizuoja savo portalų aktyvumą, kad sužinotų ar vartotojai naudojasi portalu taip, kaip buvo tikėtasi jį kuriant. Portalo duomenų apdorojimo ir analizės sistemų kūrimas prasidėjo gana seniai. Pirmieji bandymai buvo tinklapių serverių sisteminių įvykių žurnalų analizavimo programų kūrimas. Laikui bėgant puslapių turinys darėsi vis labiau dinamiškesnis, todėl tokiomis programomis darėsi vis sudėtingiau išgauti tikslią statistiką. Yra nemažai sukurta bei vis dar kuriamos sistemos analizuoti paprasčiausius tinklapius ar sudėtingus portalus.

2.2 Portalo duomenų apdorojimo ir analizės sistemų analogai

Portalo duomenų apdorojimo ir analizės sistemas pagal savo veikimo principą galima skirstyti į 3 grupes: integruotos į portalą, nepriklausomos nuo portalo, ir atskiros.

Integruotos į portalą, tai tokios sistemos, kurios yra neatskiriama portalo dalis. Duomenų rinkimas integruotas į paties portalo veikimo mechanizmą, duomenys yra renkami vidinėje portalo struktūroje. Tokios sistemos tinka tik konkrečiam portalui, kuriam ir buvo sukurtos.

Nepriklausomos nuo portalo veikimo, tai atskiros programos, kurios analizuoja serverio sisteminių įvykių žurnalų failus. Tokios programos atsirado anksčiausiai, yra lengviausiai įdiegiamos, tačiau patenkina ne visus analizės poreikius.

Atskirai veikianti analizės sistema yra nepriklausoma atskira sistema, kuri tam tikrų „blakių“ pagalba surenka duomenis iš portalų, po to juos analizuoja. Tokių sistemų diegimas šiek tiek sudėtingesnis, tačiau galima surinkti daug daugiau duomenų.

2.2.1 Analog 6.0

Programos[9] paskirtis generuoti lankomumo statistikos ataskaitas iš serverio sisteminių įvykių žurnalų. Programa yra nemokama, platinama pagal GNU/GPL licenzija. Palaiko beveik visas pagrindines operacines sistemas. Naudoja sisteminius įvykių žurnalus esančius tekstiniame formate. Diegiant tereikia nurodyti sisteminių įvykių žurnalų failų formatą. Generuojamos ataskaitos yra html formate. Analizuojama paprasčiausia bendra lankomumo statistika, priklausomai nuo laikotarpio, dažniausiai ieškomo žodžio frazė, operacinių sistemų ir naršyklių naudojimas. Tai praktiškai etaloninė tinklapių serverio sisteminių įvykių žurnalų analizės programa, kuri pradėta kurti viena iš pirmųjų. Ši programa labiausiai tinka statiškai realizuotum interneto svetainėms, kur nėra didelis apkrautumas.

2.2.2 AWStats 6.5

Programos[10] paskirtis generuoti lankomumo statistikos ataskaitas iš serverio sisteminių įvykių žurnalų. Programa yra nemokama, platinama pagal GNU/GPL licenzija. Palaiko daugumą operacinių sistemų. Yra skirta konkretiems tinklapių serverių sisteminių žurnalų failų formatams. Duomenų statistika atnaujinama kaskart peržiūrint ataskaitas. Sistema yra realizuot *Perl* programavimo kalba, todėl diegiant, jos palaikymas yra būtinas. Ataskaitos generuojamos html formatu. Ši programa generuoja daug daugiau ataskaitų, negu prieš tai apžvelgta *Analog*. Analizuojama informacija apie apsilankymus puslapiuose, naršyklių, operacinių sistemų, lankytojų ekrano rezoliucijos informacija. Šioje programoje yra nustatomas vartotojo apsilankymas, ir sekama apsilankymo trukmė. Taip pat šios programos pagalba galima sekti, robotų apsilankymus. Kad gauti kai kurią informaciją papildomai reikia modifikuoti indeksinius svetainės failus. Ši analizės sistema turi pakankamai galimybių, analizės duomenys yra kaupiami į atskirą duomenų bazę, todėl peržiūrint duomenis yra analizuojami tik dar neišanalizuoti duomenys ir jais analizės metu yra papildoma duomenų bazė.

2.2.3 phpOpenTracker 1.5.2

Ši sistema[11] renka duomenis įterpiant „blakę“ į kiekvieną puslapį. Yra nemokama ir platinama pagal GNU/GPL licencija. Programos veikimui yra reikalingas tinklapių serveris palaikantis PHP scenarijų kalbą. Jei portalas yra suprogramuotas PHP, tada galima įterpti į

puslapius PHP kodo gabalą, jei kita programavimo kalba, tada reikia įterpti HTML kodo dalį. Ši sistema turi atskirą savo duomenų bazę, į kurią yra renkami duomenys iš portalo. Papildomai, į surinkimo mechanizmą, gali būti perduodami specifiniai konkrečioms užduotims reikalingi duomenys. Ši sistema gali būti įdiegta fiziškai kitame serveryje, negu analizuojamas tinklapis. Šioje sistemoje nėra sukurtų priemonių peržiūrėti ataskaitoms, yra pateikiama tik PHP biblioteka, su tam tikrais metodais, kurių pagalba galima generuoti daugumą ataskaitų, įskaitant ir *click-path* analizę. Todėl norit naudotis šia sistema reikia papildomai susiprogramuoti peržiūros priemones naudojantis PHP scenarijų kalba. Portalo ataskaitos yra generuojamos realiame laike, tai leidžia padaryti tai, kad duomenys yra surenkami į pritaikytas tam reikalui duomenų bazės struktūras. Nedidelio lankomumo tinklapiuose ši sistema gana efektyvi, tačiau augant apkrautumui padidėja sistemos resursų naudojimas.

2.2.4 ZoomStats 1.0.2

Ši sistema[12] panaši į prieš tai aptarta phpOenTracker. Ji yra taip pat nemokama, bei platinama pagal GNU/GPL licenzija. Sistemos duomenų surinkimui reikia į kiekvieną puslapį įterpti po dalį kodo. Sistema yra realizuota taip pat PHP scenarijų kalbos pagrindu. Sistemos veikimui reikalingas tinklapių serveris palaikantis PHP. Sistema turi atskirą savo duomenų bazę, tačiau palaiko tik MySQL duomenų bazę. Ši sistema turi parengtą daugumą standartinių ataskaitų generavimą.

2.2.5 ClickTracks Analyzer

ClickTracks[13] kompanija siūlo netgi keletą produktų tinklapių analizei priklausomai nuo poreikių. Jų sistema *ClickTracks Analyzer* galima naudoti dvejopai. Vienas būdas nusipirkti visą produktą už 549 dolerius, ir pasileisti serverį pas save taip saugant analizės duomenis pas save ir išvengiant atskiro analizės serverio palaikymo. Kitas variantas yra pirkti analizės paslaugą. Kaina priklausomai nuo lankytojų srauto prasideda nuo 49 dolerių mėnesiui. Analizei ir ataskaitų generavimui naudojama atskira programa. Duomenys gali būti imami iš serverio sisteminių įvykių žurnalų, ar iš surinktų duomenų pasitelkiant į puslapį įterpto JavaScript kodo dalį. Sistema generuoja daugumą lankomumą ataskaitų, taip pat turi unikalias „sluoksnines“ ataskaitas, kurios uždeda vizualų sluoksnį ant puslapio, kuriame parodomi skaičiai, kiek kartų vartotojai paspaudė ant tam tikros vietos.

2.2.6 Top100

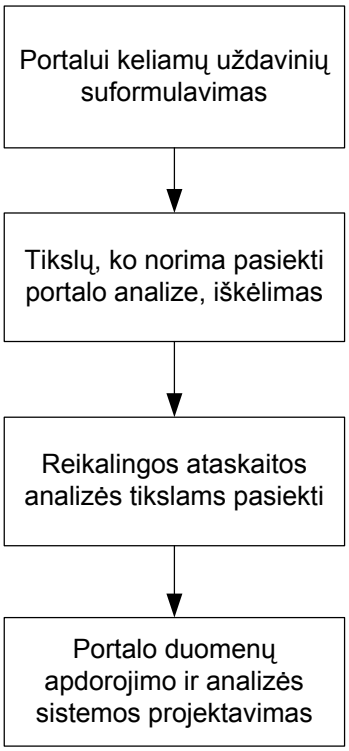
Top100[14] viena populiariausių analizės sistemų Lietuvoje. Sistema teikia tik „on-line“ paslaugas. Informacija surenkama įterpiant kodo dalį į puslapį. Ataskaitos peržiūrimos naršyklės pagalba. Teikiamos paslaugos yra ir mokamos, ir nemokamos. Už papildomą mokestį yra pateikiama daugiau ataskaitų. Taip pat yra reitinguojami visi šios sistemos naudotojai, todėl šios sistemos pagalba galima pritraukti naujų lankytojų. Ši sistema visiškai netinka intranete naudojamo portalo analizei.

2.2.7 Panašių sistemų apžvalgos apibendrinimas

Be visų čia apžvelgtų sistemų yra daug ir kitų. Dauguma stambių kompanijų išleistų analizė sistemų turi unikalių savybių, tačiau jos visus sukurtos naudojantis tam tikra metodika. Yra naudojami du pagrindiniai duomenų šaltiniai: tinklapių serverių sisteminiai įvykių žurnalai, ir įterptos „blakės“ į tinklapių kodą. Didžiausia problema tokių sistemų yra duomenų integracija tarp portalo ir analizės duomenų, todėl norint šias pritaikyti saviems poreikiams reikia papildomai į tai investuoti. Nemokamos sistemos turi savų trūkumų, jose nėra visų reikalingų ataskaitų, be to jose sunkiai galima realizuoti duomenų integracija tarp portalo ir analizės duomenų bazių.

2.3 Tinklapių analizės metodika ir technologijos

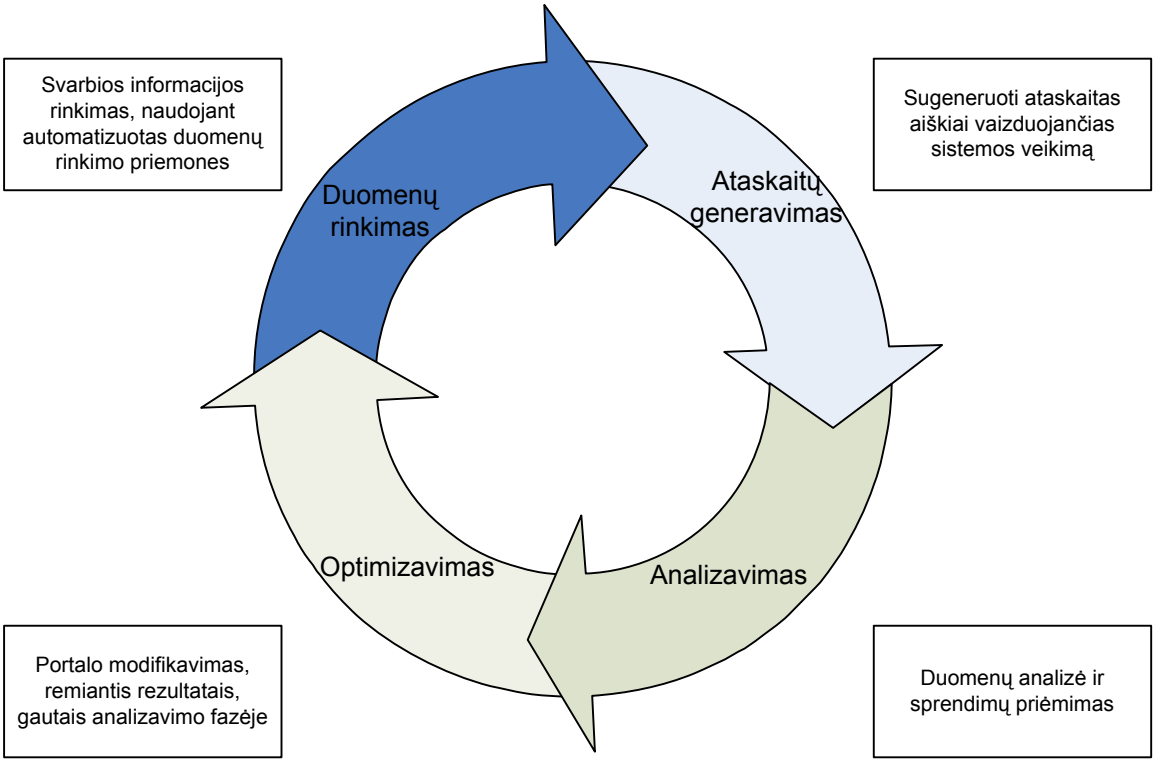
Kiekvienas portalas yra unikalus, todėl pritaikyti kažkokią jau sukurtą sistemą galima tik iš dalies. Projektuojant naują portalą po kiek laiko atsiranda analizės įrankių poreikis. Tada iškyla klausimas ar kurti savo įrankius ar pritaikyti jau esamus savo reikmėms. Tačiau netgi norint naudoti jau sukurtą analizės sistemą savo poreikiams, reikia suprasti jų veikimo principą, bei atlikti jų galimybių studiją, kad pasirinkti sistemą geriausiai atitinkančią poreikius.



1 pav. Portalo analizės sistemos projektavimas

Portalo analizės sistemos kūrimas prasideda ne nuo techninės dalies. Prieš pradedant tokios sistemos projektavimą, reikia išsiaiškinti portalo paskirtį, vartotojų poreikius, uždavinius iškeltus portalui. Tada išsikelti tikslus, ką norima pasiekti portalo analize. Turint suformuotus analizės keliamus uždavinius galima pradėti ieškoti kokių reiktų ataskaitų, kad jos atsakytų į iškeltus klausimus. Pagal sugeneruotų ataskaitų duomenis atitinkamai modifikavus portalą ar įtakojus portalo vartotojus, turėtų būti siekiami analizės tikslai. Žinant kokių norima ataskaitų iš portalo, galima pradėti projektuoti portalo duomenų apdorojimo ir analizės sistemą.

Analizės sistemoje yra dvi atskiros dalys, kurios gali veikti nepriklausomai viena nuo kitos, tai duomenų surinkimo ir analizavimo, ataskaitų generavimo. Surinkus informacija, kaip lankytojai elgiasi portale, bei ją išanalizavus, galima toliau tobulinti portalą, vystyti vartotojo sąsają. Duomenų kaupimas ir jų analizavimas be jokio portalo įtakojimo ateityje yra beprasimis pajėgumų švaistymas. Visą analizės ir portalo tobulinimo eigą galima įsivaizduoti kaip niekada nenutrūkstantį procesą [1] (2 pav.).



2 pav. Nenutrūkstantis portalo tobulinimo procesas

Visų pirma yra surenkami duomenys apie portalo aktyvumą. Kita fazė yra iš surinktų duomenų sugeneruoti ataskaitas. Toliau žmonės, priimantys sprendimus, analizuoja duomenis ir sprendžia, kaip reiktų modifikuoti portalą, kad būtų padidintas jo efektyvumas arba, kaip daryti įtaką vartotojams, kad pasiteisintų lūkesčiai, siejami su portalu. Kai kurie autoriai įžvelgia ir daugiau fazių šiame portalo tobulinimo procese. Tačiau šios yra pagrindinės ir svarbiausios.

2.4 Duomenų rinkimo šaltiniai

Prieš pradedant portalo duomenų analizę, būtina žinoti, iš kur tie duomenys gali būti renkami, bei kokius duomenis galima rinkti. Ką galima analizuoti portale? Atsakant į šį klausimą reikia suprasti kokie duomenų srautai yra portaluose. Šie duomenys pagal savo kilmę yra skirstomi į kelias grupes[3]:

- Tinklo srauto duomenys (*Web traffic data*) – tai seka, kuria lankytojas kreipėsi į tinklapių serverio informaciją. Tai pats populiariausias ir svarbiausias duomenų šaltinis. Dažniausiai šie duomenys „iškasami“ iš tinklapių serverio sisteminių žurnalų failų, taip pat yra sukuriami „blakių“ įterptų į puslapius. Ši informacija

yra formuojami programos, kuri daro įrašus kiekvienos užklausos metu (pvz. paspaudžiant ant nuorodos).

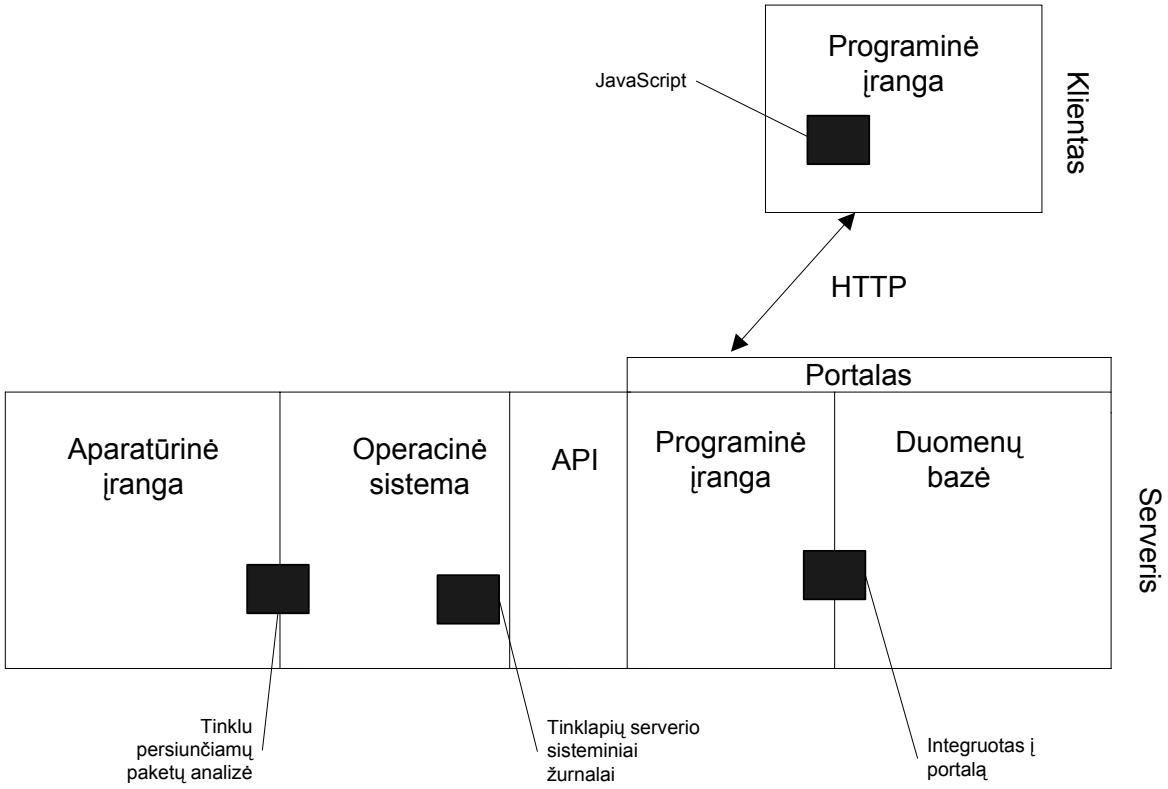
- Portalo transakcijos duomenys *(Web transactional data)* – šie duomenys panašūs į prieš tai aprašytus tinklo srauto duomenis. Jie labiau skirti elektroninės komercijos analizei. Ši informacija apima: klientų skaičių, užsakymų skaičių bei vidutinį transakcijos didį.

- Tinklapių serverio našumo duomenys *(Web server performance data)* – šie duomenys apima, laiką per kiek yra užkrautas puslapis, koks duomenų kiekis yra persiunčiamas puslapiui užkrauti, kiek ir kokių įvyko klaidų generuojant puslapį.

- Naudojimo studijos *(Usability studies)* – šios studijos yra labiau atskiras mokslas, kuris įtraukia darbą su tikrais žmonėmis, kad geriau suprasti kaip jie naudojasi portalu.

- Vartotojo persiųsti informacija ir panašūs duomenys *(User submitted information and related data)* – šie duomenys apima informaciją surinktą naudojant apklausas, tiesioginius vartotojų duomenų įvedimus į formas ar tiesiai į duomenų bazę.

Literatūroje dažniausiai mini du duomenų surinkimo būdai kaip pagrindiniai [3,2] – tai duomenų rinkimas naudojant sisteminius įvykių žurnalus, bei „balkių“ įterpimas į kiekvieną puslapį. Taip pat mažai paplitęs, tačiau galingas duomenų rinkimo būdas yra analizuoti tinklu siunčiamus paketus. Tačiau praktikoje galima sutikti dar vieną duomenų rinkimo būdą, kai visas mechanizmas yra integruotas į portalą. Bet kuris duomenų rinkimo būdas turi savų privalumų ir trūkumų, todėl norint pasiekti geriausių rezultatų, galima naudoti kelis duomenų šaltinius. Kiekvienas duomenų šaltinis yra skirtingoje terpėje, todėl jų pagalba yra gaunami skirtingi duomenys



3 pav. Duomenų šaltiniai

3 pav. pavaizduota, kaip yra išsidėstę pagrindiniai duomenų šaltiniai bendrame programinės ir aparatūrinės įrangos požiūriu.

2.4.1 Tinklapių serverio sisteminiai žurnalai *(Web server log files)*

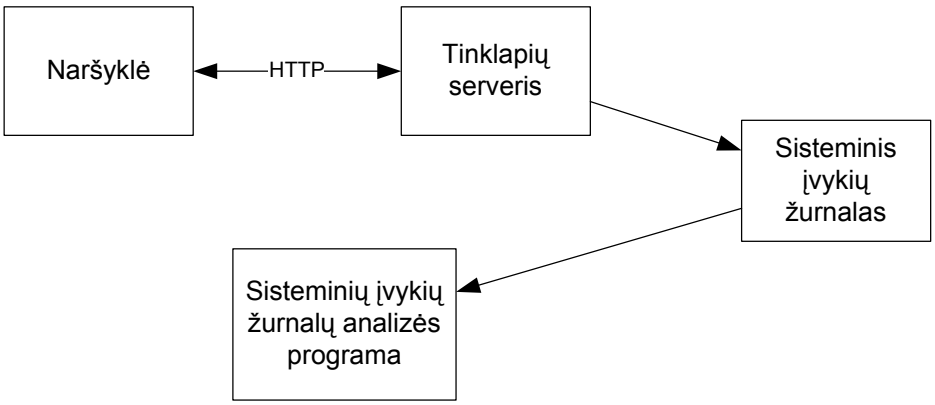
Anksčiausiai atsiradęs ir labiausiai paplitęs analizės metodas tai tinklapių serverio sisteminių žurnalų failų analizavimas. Jo paplitimo priežastis – lengvas įdiegimas. Dažniausiai, paleidus tinklapių serverį su standartine, konfigūracija būna jau veikiantis įvykių registravimas į serverio sisteminių žurnalų failą. Todėl nesunkiai galima grįžti prie įvykių, kurie buvo užregistruoti, dar prieš numatant naudoti duomenų analizavimo galimybę.

Tinklapių serverio paskirtis – pateikti resursus jų užklausiantiems vartotojams. Resursai yra objektai tokie, kaip HTML puslapiai, paveiksliukai, daugialypės terpės failai, scenarijų failai ir panašiai. Užklausimus dažniausiai siunčia naršyklės naudojamos žmonių, tačiau tai gali būti ir automatiniai agentai ar vorai indeksuojantys puslapius. Tinklapių serveris veikia gerai tol, kol gali patenkinti visas užklausas. Tokių užklausų sekimas ir yra rašomas serverio sisteminius įvykių žurnalus.

Tinklapių serverių sisteminiai žurnalai dažniausiai yra tekstiniai failai, kai kuriais atvejais jie gali būti xml formatu, ar rašomi tiesiai į duomenų bazę. Skirtingi tinklapių serveriai siūlo skirtingus sisteminių žurnalų failų formatus.

2.4.1.1 Duomenų rinkimo architektūra

Duomenų rinkimo mechanizmas yra integruotas tinklapių serveryje. Šį mechanizmą galima išjungti arba įjungti, pakeičiant tinklapių serverio konfigūraciją. Dauguma šiuolaikinių tinklapių serverių turi galimybę, rašyti sisteminius įvykių žurnalus. 4 pav. pavaizduoti duomenų srautai naudojant sisteminius įvykių žurnalus kaip duomenų šaltinį portalo.



4 pav. Duomenų srautai naudojantis sisteminiais įvykių žurnalais

Naudojant tokią architektūra duomenų surinkimo sistema gali funkcionuoti, nepriklausomai nuo analizės sistemos. Kadangi duomenys rašomi į sisteminius įvykių žurnalus yra standartizuoti, tai galima naudoti kelias analizės programas. Naudojant tokį duomenų šaltinį, analizės programos būna kelių tipų. Vienos kiekvieną kart generuojant ataskaitos analizuoja visą įvykių žurnalą, kitos turi savo vidinę duomenų bazę, ir prieš generuojant ataskaitas analizuoja tik papildytą, nuo paskutinės analizės, dalį.

2.4.1.2 Galimi duomenys

Priklausomai nuo tinklapių serverio sisteminiai įvykių žurnalai yra tam tikro formato. Tipiškas įrašas sisteminiame įvykių žurnale CLF formate atrodo taip:

```
127.0.0.1 - frank [10/Oct/2005:13:55:36 -0700] "GET /apache_pb.gif
HTTP/1.0" 200 2326
```

Yra įvairių priedų prie tinklapių serverių, kuriuos įdiegus galima išplėsti informaciją, registruojamą įvykių žurnale.

1 lentelė. Informacija rašoma į sisteminius įvykių žurnalus

Įrašo savybė	Paaiškinimas
Užklaustas resursas	Per HTTP užklaustas failas (HTML, GIF, JPG, PDF ir pan.)
Data	Užklausos data paremta serverio data
Laikas	Užklausos laikas paremtas serverio laiku
Kliento IP adresas	IP adresas naršyklės kuri daro užklausą
Nuodytojas	URL turintis nuorodą į resursą iš kurios buvo ateita į puslapį
Vartotojo agentas	Naršyklė ar programa, kuri siunčia užklausą
Paslaugos pavadinimas	Tinklo programos pavadinimas
Serverio pavadinimas	Serverio aptarnaujančio resursą pavadinimas
Serverio IP adresas	Serverio aptarnaujančio resursą IP adresas
Metodas	GET arba POST. Šie abu metodai yra informacijos persiuntimui tarp serverio ir kliento.
URI užklausa	Užklausos eilutė pridėta prie galo URL. Ne visada bet dažniausiai naudojama perduoti parametrams dinamiškai generuojamiems puslapiams.
HTTP statusas	Skaitinė reikšmė aprašanti tinklapio serverio atsakymą. Visas sąrašas gali būti rastas http://www.w3.org/Protocols/rfc2616/rfc2616-sec6.html
Baitu išsiusta	Bendras kiekis baitu persiustas tinklapio serverio, kad patenkinti resursų užklausimą.
Baitu gauta	Bendras gautas baitu kiekis tinklapio serverio, kad patenkinti resurso užklausą.
Užtrukta laiko	Bendras laiko kiekis per kurį užklaustas ir persiųstas resursas
Sausainėlis	Tekstas kuris turėjo būti įrašytas į sausainėlį.

Išplėsto formato tinklapio serverio sisteminio įvykių žurnalo failo formatas(CLF):

```
lothlorien.ncsa.uiuc.edu - - [19/Sep/1995:15:19:07 -0500] "GET
/images/icon.gif HTTP/1.0" 200 1656 "http://hoohoo.ncsa.uiuc.edu/"
"NCSA_Mosaic/2.7b1 (X11;IRIX 5.3 IP22) libwww/2.12 modified"
```

Microsoft IIS tinklapių serverio failo formatas:

```
172.16.255.255,anonymous,03/20/98,23:58:11,MSFTPSVC,SALES1,192.168.114
.201,60,275,0,0,0,PASS,intro.htm
```

Kaip matoma iš pavyzdžių, skirtingi tinklapių serveriai renka skirtingą informaciją, bei skirtingais formatais. Sisteminių žurnalų failuose yra kaupiamos užklausos ne tik apie puslapių, bet taip pat ir paveiksliukų, bei kitų rinkmenų reikalingų užkrauti visą puslapį. Esant dideliam vartotojų skaičiui, užklausų skaičius gali būti nemažas, todėl įvykių žurnalų

dydis auga greitai, žurnalų failai užima nemažai vietos bei reikia nemažai resursų juos išanalizuoti.

2.4.1.3 Privalumai

- Duomenų nuosavybė – visi sukurti sisteminiai įvykių žurnalų failai yra sukuriami tuose pačiuose serveriuose kur ir tinklapis.
- Duomenų surinkimo lankstumas – nesunkiai galima pakeisti surenkamų duomenų formatą, kad surinkti tik tuos duomenis kurie reikalingi analizei.
- Lengvas įdiegimas – įdiegiant nereikia modifikuoti kiekvieno tinklapio puslapio, kadangi tai yra tinklapių serverio realizacijos dalis.
- Galimybė rašyti tiesiai į duomenų bazę – tai palengvina analizės programų kūrimą, nes nereikia kurti savo duomenų surinkimo šaltinio rašančio duomenis į duomenų bazę.
- Galimybė sekti ar užklausos baigėsi sėkmingai. Tinklapių serverių sisteminiai įvykių žurnalų failai seka persiųstų duomenų kiekį, bei HTTP statusą, kurių pagalba galima nustatyti ar užklausos buvo įvykdytos sėkmingai ir ar visas failas buvo persiųstas.
- Galimybė sekti robotų ir vorų apsilankymus tinklapyje, kadangi yra registruojamas naršančio agento vardas.

2.4.1.4 Trūkumai

- Greitai auganti užimama disko vieta – esant dideliam vartotojų lankomumui (pvz 1000 ir daugiau vartotojų per dieną), sisteminių įvykių žurnalų failų užimama vieta diske gali stipriai išaugti.
- Kešavimas (*Cache*) – naudojantis internetu tiek naršyklės tiek tarpiniai (*proxy*) serveriai saugo atsargines puslapių kopijas, todėl kreipiantis vartotojui į puslapį ne visada yra kreipiamasi į serverį, todėl šios užklausos nebūna įrašomos į sisteminius įvykių žurnalus.
- IP adresas kaip unikalus vartotojo identifikatorius - tai gaunasi sudėtingas uždavinys, kadangi vieną lankytoją bandoma identifikuoti pagal jo tinklo adresą, žinant, kad kai kurių lankytojų tinklo adresas gali dažniai keistis, bei gali būti keli lankytojai, turintys tą patį adresą.
- Atskirų puslapių identifikavimas - įvykių žurnaluose registruojami kreipiniai į tam tikrus failus. Portale vienas vizualiai matomas puslapis gali būti sudarytas iš kelių failų, ar skirtingi puslapiai, matomi kreipiantis į tą patį failą su skirtingais

parametrais. Tokiu atveju sunku nustatyti, į kurią loginę portalo dalį buvo kreiptasi.

- Ataskaitų generavimo laikas – kadangi sisteminiai įvykių žurnalai gali greitai išaugti, tai jų analizavimas užima taip pat nemažai laiko.

2.4.2 Duomenų rinkimas naudojant „blakės“ technologiją

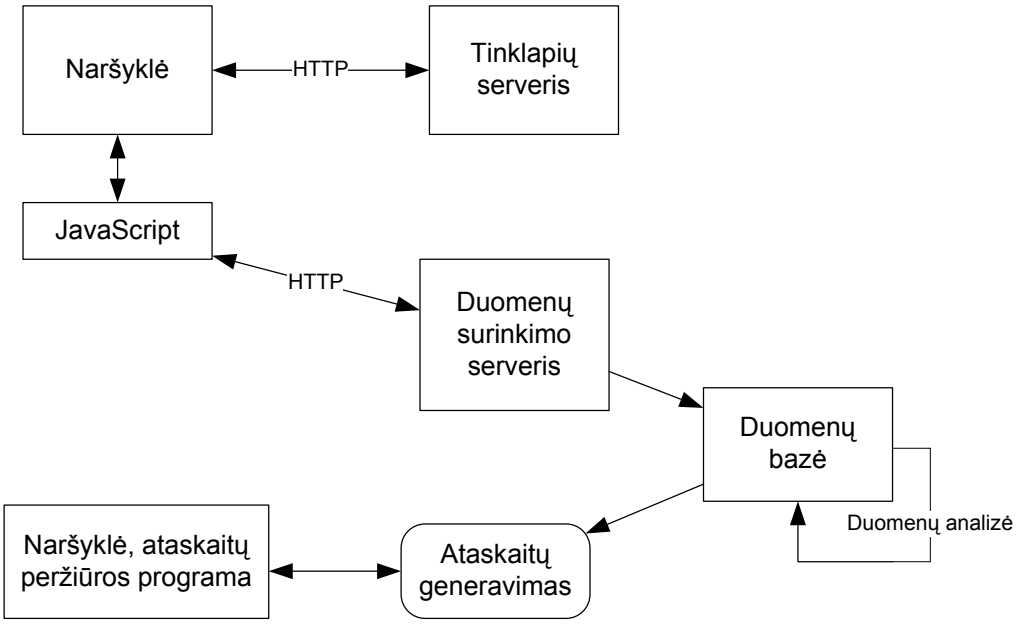
Vienas iš labiausiai pastaruoju metu paplitusių būdų, yra įterpti „blakę“ į kiekvieną portalo puslapį, ir taip surinkti labai svarbią informaciją apie peržiūrimą puslapį ir apie vartotoją, peržiūrintį puslapį. Kas tai per „blakė“? Į kiekvieną puslapį yra įterpiamas kodo dalis, tai gali būti kodas kuri interpretuojamas serveryje, bet dažniausiai būna interpretuojamas kliento naršyklėje. Pats populiariausias yra įterpti JavaScript kodo dalį, todėl kad tokiu būdu yra įmanoma surinkti daugiau duomenų apie lankytoją. JavaScript dinamiškai sugeneruoja išorinio serverio užklausą, sudėdamas visą surinktą informaciją į užklausos eilutę. Taip surinkti duomenys yra persiunčiami į analizės serverį.

Šio duomenų rinkimo metodo populiarumą lėmė tai, kad atskiros kompanijos gali teikti komercines paslaugas. Jos pas save turi serverį, kuris atlieka analizės funkcijas, klientui tereikia į savo portalo kiekvieną puslapį įterpti duotą JavaScript kodą. Tada duomenys yra renkami paslaugą teikiančios kompanijos serveryje, ten yra taip pat priemonės, generuoti ataskaitoms. Internete yra nemažas tokias paslaugas teikiančių kompanijų skaičius, siūlančių įvairias analizės priemones.

Tokią duomenų surinkimo schemą galima naudoti ir intranete naudojamam portalui, turint savo surinkimo ir analizės serverį, tada išsprendžiama problema kuri iškyla dėl duomenų nuosavybės.

2.4.2.1 Duomenų rinkimo architektūra

5 pav. pavaizduoti duomenų srautai įterpiant JavaScript blakę ir naudojant tam tikrą duomenų analizės paslaugų tiekėją.



5 pav. Sistemos architektūra naudojantis teikiamomis analizės paslaugomis

Naudojant tokią architektūrą būtinai reikalingas internetas. Jei dingsta ryšys su paslaugas teikiančiu serveriu, tai duomenys yra nesurenkami. Taip pat būtina, kad klientas peržiūrintis portalą, turėtų internetinį ryšį ir su analizės serveriu. Šiuo atveju analizės sistema yra visiškai nepriklausoma nuo portalo.

Paslaugų teikėjų duomenų saugojimo formatas dažniausiai yra nežinimas, bet duomenys yra saugomi dažniausiai duomenų bazėse. Tokiu būdu gali būti surenkama įvairi informacija, taip pat kaip ir naudojant sisteminius įvykių žurnalus, tačiau duomenų saugojimo formatas yra pasirenkamas sistemos projektuotojų.

2.4.2.2 Privalumai

- Tikslumas – kadangi duomenys yra surenkami galutinio vartotojo naršyklėje, o ne tinklapių serveryje, todėl duomenys apie vartotoją yra daug tikslesni.
- Ataskaitų generavimo laikas – surenkant duomenis galima juos iškart transformuoti į duomenų struktūras reikalingas greitam ataskaitų generavimui.
- Duomenų lankstumas – šis metodas priklauso nuo paslaugos tiekėjo duomenų formato, tačiau leidžia rinkti apibrėžtus duomenis kaip vartotojo ID, grupės ID ir pan.

2.4.2.3 Trūkumai

- Reikalingas internetas – tai pagrindinis trūkumas naudojant išorinį duomenų surinkėją.

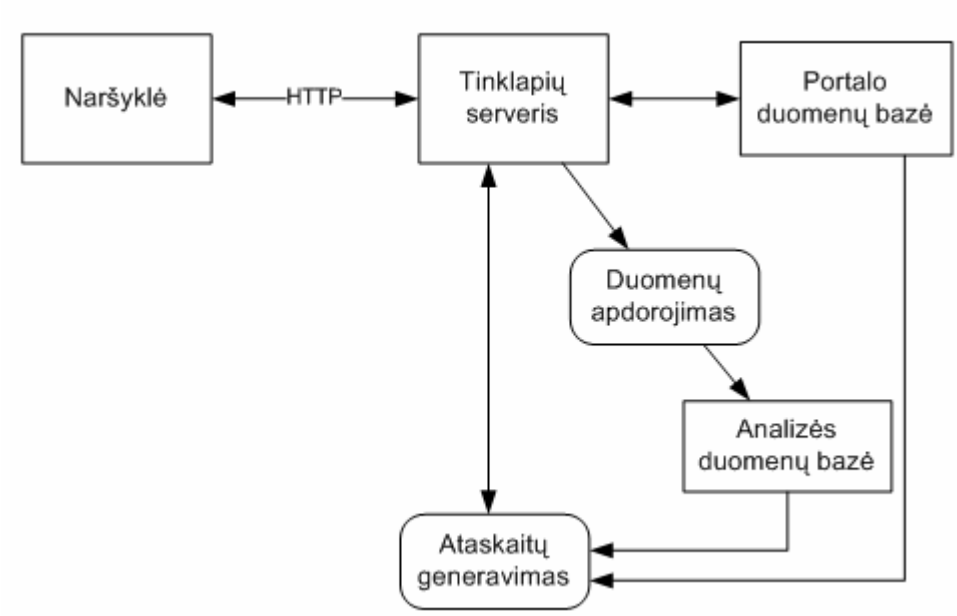
- Priklausomybė nuo JavaScript - vartotojo naršyklė būtinai turi palaikyti JavaScript, kad būtų galima rinkti informaciją.
- Priklausomybė nuo sausainėlių, kaip unikalaus vartotojų identifikatoriaus – pas vartotojus turi būti leista rašyti sausainėlius, bei kitos su tuo susijusios problemos.
- Sudėtingas diegimas - minusas tokio duomenų surinkimo modelio tas, kad svarbu sudėti „blakes“ į kiekvieną puslapį, jei jų nėra duomenys nėra surenkami. Todėl diegimo klaidos gali atsiliepti duomenų tikslumui.
- Našumo problemos – kiekviena papildoma kodo eilutė reikalauja papildomų kompiuterinių resursų, todėl bet kokio kodo papildymas sąlygoja ilgesnį puslapių krovimo laiką.
- Automatinių robotų ir vorų – šios sistemos pagalba negalima užfiksuoti, kadangi jie nevykdo JavaScript kodo. Bet tai gali būti palaikytas kaip ir privalumas, nes šios automatinės programos neiškreipia duomenų.
- Privatumo užtikrinimas – jei naudojamas paslaugų tiekėjas, tai duomenys yra saugomi pas jį, todėl neaišku, kaip šiuos duomenis kas nors gali panaudoti.

2.4.3 Duomenų rinkimas integruotas į portalą

Vienas iš būdų padedantis rinkti duomenis yra galimybė visą duomenų rinkimo procesą integruoti į portalą. Tai panašu į tinklapio serverio sisteminių žurnalų generavimą ar netgi „blakių“ įterpimą, skirtumas toks, kad galima rinkti tik reikalingus duomenis, atsižvelgiant į konkrečią portalo loginę architektūrą. Tokios sistemos yra jau sukurto portalo dalis, todėl kai kurie laisvai prieinami ar komerciniai portalai turi savyje integruotas duomenų rinkimo ir analizės sistemas. Nėra siūloma tokiu principu jau realizuotų portalų analizės sistemų, kurias būtų galima naudoti savo poreikiams, todėl dažniausiai portalų kūrėjai turi patys organizuoti tokį duomenų surinkimą. Pagrindinis tokio mechanizmo trūkumas, kad sukurtą duomenų surinkimo ir analizės sistemą dažniausiai sunku panaudoti kitam portalui.

2.4.3.1 Duomenų rinkimo architektūra

Toks duomenų rinkimo būdas, leidžia duomenų rinkimo metu praktiškai iškart apdoroti duomenis, aišku tai papildomai kainuoja sistemos resursų, tačiau kai kuriais atvejais tikslus ir greitas ataskaitų generavimas yra labai svarbus. Taip surinkti duomenys gali būti nesunkiai klasifikuojami pagal vidinę portalo struktūrą, kas labai pagerina analizės galimybes. Tai pat nereikia atskiro duomenų rinkimo serverio, tam yra naudojamas tas pats, kurį naudoja ir portalas.



6 pav. Duomenų rinkimo mechanizmas, integruotas į portalą

Duomenų srautai naudojant į portalą integruotą duomenų šaltinį pavaizduoti 7 pav. Tas pats tinklapių serveris kiekvienos užklausos metu apdoroja ir statistinius duomenis įrašydamas į atskiras analizei reikalingas duomenų bazės lenteles. Tokiu būdu statistikos duomenys yra iškart apdorojami, ir ataskaitas galima iškart peržiūrėti. Tačiau jei norima tokiu būdu generuoti skirtingų tipų ataskaitas, gali tekti rašyti tuo pat metu į daug skirtingų lentelių, kas apsunkintų tokios sistemos projektavimą ir darbą.

2.4.3.2 Privalumai

- Lankytojų naršyklės nustatymai nedaro įtakos sistemos veiklai.
- Ataskaitų generavimo laikas – surenkant duomenis galima juos iškart transformuoti į duomenų struktūras reikalingas greitam ataskaitų generavimui.
- Duomenų nuosavybė – analizės duomenys yra neatsiejami su portalo duomenimis.
- Galima portalo vidinės struktūros analizė, lengvai galima identifikuoti konkrečių vartotojų priėjimą prie resursų (straipsnių, straipsnių grupių ir pan.)

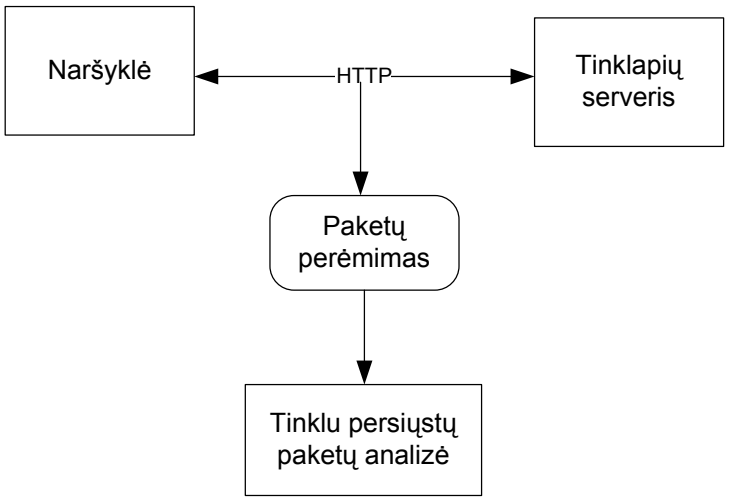
2.4.3.3 Trūkumai

- Sudėtingas diegimas ir modifikavimas, kadangi ši sistema yra integruota į portalą tai jos diegimas yra neatsiejama portalo projektavimo dalis.
- Nėra galimybės atskirti robotų indeksuojančių puslapius naršančių portalą.

- Našumo problemos – kiekviena papildoma kodo eilutė reikalauja papildomų kompiuterinių resursų, todėl bet kokio kodo papildymas sąlygoja ilgesnį puslapių krovimo laiką.

2.4.4 Tinklu siunčiamų paketų analizavimas

Tinklu siunčiamų paketų analizė (*Sniffing*), taip pat yra vienas iš galimų duomenų rinkimo modelių. Privalumas tokio šaltinio tas, kad galima surinkti daug techninės informacijos, tokios kaip atšauktos ar nutrūkusios užklausos, duomenų srauto apkrautumas. Pagrindinis privalumas tokio duomenų rinkimo yra tas, kad galima gauti tokius duomenis, per kiek laiko vartotoją pasiekia užklausą, ar iš vis pasiekia. Toks analizavimas skirtas išsiaiškinti portalo našumą tinkle. Tokiu būdu yra praktiškai neįmanoma stebėti vidinės portalo struktūros. Tokios duomenų analizavimo schemos duomenų srautai pavaizduoti 7 pav.



7 pav. Duomenų srautai naudojant tinklu siunčiamų paketų analizę

Tinklu siunčiamus paketus perėmus, jų antraštės rašomos į duomenų bazę, ar įvykių žurnalą. Žinant koks paketų skaičius persiunčiamas tinklu, įrašų gali kauptis tikrai daug, todėl pastovus tokiu būdu informacijos kaupimas dažniausiai nenaudojamas.

2.4.5 Duomenų rinkimo šaltinių palyginimas

Kiekvienas iš išvardintų šaltinių turi savų privalumų ir trūkumų. 2 lentelėje yra lyginamos duomenų šaltinių savybės, į kurias reikia atsižvelgti renkantis duomenų rinkimo šaltinį.

2 lentelė. Duomenų surinkimo šaltinių apžvalga

Poreikiai, apribojimai, savybės	Įvykių žurnalai	Paketų analizė	JavaScript kodas	Integruotas
Renkami duomenys prieš analizės sistemos įdiegimą	Taip	Ne	Ne	Ne
Poreikis įdiegiant modifikuoti kiekvieną puslapį	Ne	Ne	Taip	Taip
Naršyklių saugyklos ir automatiniai naršymo agentai daro įtaką duomenų tikslumui	Taip	Taip	Ne	Ne
Nereikalinga papildoma aparatūrinė įranga	Ne	Taip	Ne	Ne
Priėjimas prie informacijos realiame laike	Ne	Galima realizuoti	Galima realizuoti	Galima realizuoti
Reikalinga serverio apkrautumo informacija	Taip	Ne	Ne	Taip
Informacijos nuosavybė	Taip	Taip	Jei išorinis ne	Taip
Skirtingų rinkmenų tipų sekimas	Taip	Ne	Ne	Taip
Įdiegimo laikas ir lengvumas	Paprasčiausias	Sudėtinga	Vidutinis	Sudėtinga
Duomenų rašymas tiesiai į duomenų bazę	Taip	Taip	Per tarpinį serverį	Taip
Analizės priemonių kaina ir pasirinkimas	Didelis	Mažai	Didelis	Nėra
Daro įtaką naršyklės nustatymai	Ne	Ne	Taip	Ne

Esant poreikiui galima naudoti kelis ar netgi visus duomenų šaltinius, taip surinkti visus įmanomus duomenis. Naudojant kai kuriuos duomenų šaltinius kartu galima išgauti duomenis, kurių neįmanoma gauti naudojant bet kurį iš jų atskirai.

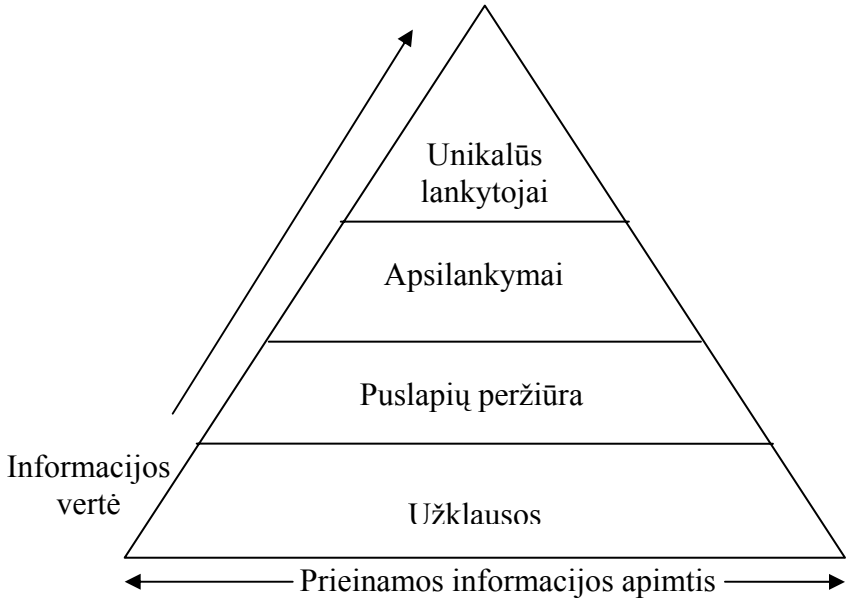
2.5 Analizuojama informacija portale

Portalas sudėtinga sistema, todėl prieš apsisprendžiant, ką analizuoti, reikia žinoti, koks yra visos analizės tikslas. Analizės metu galima bandyti išsiaiškinti, kiek lankytojų portalas sulaukia per dieną, kurie puslapiai populiariausi. Jei tai vidinio tinklo portalas, galima analizuoti ar tie vartotojai, kurie privalėtų naudotis, tikrai naudojasi portalu. Visų pirma reikia iškirti kurie portalo naudojimo rodikliai yra svarbūs atliekamai analizei. Reiktų įsivesti kintamuosius, kuriuos galima išmatuoti. Dažniausiai tai būna lankytojai, puslapiai, resursai, lankymosi laikas, puslapiai iš kur ateina nauji vartotojai. Galima analizuoti daug daugiau

aspektų, tačiau lankytojai ir puslapiai yra pagrindiniai kriterijai, kurie yra svarbiausi portalo analizėje.

2.5.1 Lankytojų klasifikacija

Lankytojai yra pats svarbiausias portalo analizavimo atributas. Lankytojas apibrėžiamas kaip žmogus ar automatinė programa naršantys portale. 8 pav. rodo, koks yra duomenų svarbumas iš visų surinktų duomenų apie lankytojus.



8 pav. Prieinamų duomenų apimtis ir svarba

Iš visos informacijos masės svarbiausia identifikuoti atskirus unikalius lankytojus, nes tai yra pagrindinis rodiklis, kuris leidžia nustatyti kiek žmonių buvo apsilankę portale. Juos galima skirstyti į kelias grupes:

1. Žinomi vartotojai – tai vartojai, turintys prisijungimo vardą, bei turima jų asmeninė informacija.
2. Dalinai žinomi vartotojai – tai vartojai, turintys prisijungimo vardą, tačiau jie nėra suteikę tikslios informacijos apie save.
3. Nežinomi vartotojai – tai vartotojai, neturintys prisijungimo vardo, juos galima identifikuoti pagal IP adresą, naršyklę ir panašiai.

Žinomų vartotojų grupė yra pati svarbiausia, nes pagal jų suteiktą asmeninę informaciją, galima vartotojus skirstyti į kategorijas, pagal tam tikrą demografinę informaciją ir taip sudaryti ataskaitas, kurios atspindėtų tam tikros grupės lankytojų aktyvumą, bei leistų priimti sprendimus norint daryti įtaką tai grupei. Vartotojų registracija labai palengvina unikalių vartotojų atpažinimą.

Tarp visų lankytojų yra galimybė stebėti ar vartotojų apsilankymai yra vienkartiniai ar vartotojai sugrįžta. Tam tikslui yra panaudojami sausainėliai (*cookies*). Dažniausiai vartotojai leidžia savo naršyklėms priimti sausainėlius, tokiu būdu kai vartotojas grįžta į portalą po pirmo apsilankymo sistema jau gali identifikuoti, kad tai yra sugrįžęs lankytojas. Įterpus unikalų sausainėlį galima atskirti unikalius lankytojus. Tačiau ši sistema yra netobula ir gali iškreipti duomenis, nes yra lankytojų kurie naudoja kelis kompiuterius, ar kompiuterių kuriais naudojasi keletas lankytojų

2.5.2 Puslapiai

Sekantis svarbus portalo resursas yra informacija, prie kurios lankytojai gali prieiti portale. Tradiciškai tai vadinama puslapiais, tačiau tas pavadinimas susiformavęs, kai didžioji dalis informacijos prieinamos HTTP protokolu buvo statiška ir puslapis atitikdavo vieną html failą. Puslapį galima apibrėžti kaip „puslapį“ informacijos prieinamą per internetą. Šiuolaikiniuose portaluose informacija dažniausiai yra generuojama dinamiškai ir dažnai pasitaiko netgi variantų, kad matomas ekrane vaizdas neturi kažkokio konkretaus URL adreso, o yra manipuliacijų vidiniais portalo kintamaisiais rezultatas. Tai pasunkina identifikuojant atskirus puslapius, kuriuos mato lankytojas. Puslapių identifikavimą gali palengvinti aiškus ir aiškiai apibrėžiantis poziciją puslapio unikalus pavadinimas. Jis yra įrašytas HTML kalboje <*title*> bloke. Blogas pavadinimas butų toks *index.html* todėl, kad tai yra failo pavadinimas, o ne puslapio pavadinimas. Geras pavadinimas turi atspindėti tai ką vartotojas mato ekrane, pvz: *Temos: Informacija apie portalus redagavimas (ID: 554566)*, toks pavadinimas aiškiai apibrėžia, kad vartotojas dabar atsidaręs puslapį, kuriame galima redaguoti temą, kurios pavadinimas „Informacija apie portalus“ ir kurios vidinis id yra „554566“.

Svarbus puslapių rodiklis yra kiek kartų jis buvo peržiūrėtas. Tai leidžia identifikuoti tam tikros informacijos populiarumą. Kitas rodiklis, tai kiek laiko lankytojai praleidžia tam tikrame puslapyje. Šis rodiklis parodo ar puslapyje informacija yra atidžiai įsisavinama ar perverčiama greitai, nekreipiant į ją per daug dėmesio. Pastaruoju metu sparčiai populiarėja *click-stream* analizė. Tai yra analizuojama seka veiksmų, kuria vartotojas nueina iki tam tikro rezultato. Šitokia analizė labai svarbi e-parduotuvėse. Sakykim analizės metu pastebime, kad 90% pirkėjų yra prarandama bandant užpildyti kokią nors sudėtingą formą, radus tokią vietą galima padaryti išvadas ir palengvinti formos pildymą, arba ją iš vis pašalinti.

2.6 Ataskaitų generavimas

2.6.1 Duomenų srautai

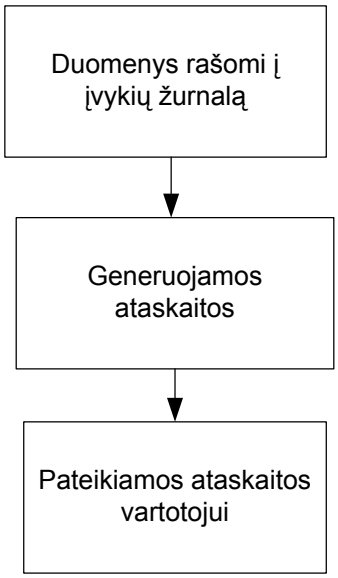
Pradedant analizuoti duomenis reikia įsivaizduoti, kokie duomenų srautai vaikšto portale. Dažniausiai visi veiksmai yra sinchroniniai vartotojas padaro veiksmą, sulaukia rezultato ir tada gali daryti sekantį veiksmą. Surenkant duomenis yra registruojama kiekviena vartotojo siunčiama užklausa. Taip yra sugeneruojamas didelis duomenų kiekis, iš kurių reikia „iškasti“ naudingą informaciją. Esant dideliam vartotojų kiekiui tokios informacijos kiekis auga labai greitai. Lankytojas būdamas portale gali peržiūrėti kelias dešimtis puslapių vieno apsilankymo metu ir kiekvienas jo veiksmas palieka įrašą duomenų bazėje ar tinklapių serverio įvykio žurnalo faile tolimesnei analizei. Prie kiekvieno veiksmo yra laikas kada šis veiksmas atliktas, bei kiek galima smulkesnis veiksmo aprašymas su visa įmanoma vartotojo informacija. Taigi rezultate yra vartotojų užklausų sąrašas kurį reikia apdoroti.

2.6.2 Svarbios informacijos atskyrimas

Iš visos sukauptos informacijos analizei nėra svarbu kada koks vartotojas ką nors darė. Svarbūs yra tik bendrai apibendrinti duomenys. Kokius puslapius vartotojai dažniausiai peržiūrėjo. Kurios vartotojų grupės aktyviausiai naudojasi portalu ir panašiai. Visa apibendrinama informacija turi padėti susidaryt bendrą vaizdą kaip yra naudojamas portalas. Taigi visą apibendrintą informaciją kažkaip reikia surinkti ir pateikti ataskaitų, lentelių ar grafikų pavidalu. Kiekvienos ataskaitos specifika reikalauja skirtingų duomenų struktūrų. Pavyzdžiui puslapių peržiūros eiliškumo ataskaita ir labiausiai lankomų puslapių ataskaita labai skiriasi savo generavimu, bei savo struktūra. Tačiau visos ataskaitos yra generuojamos iš to paties duomenų srauto.

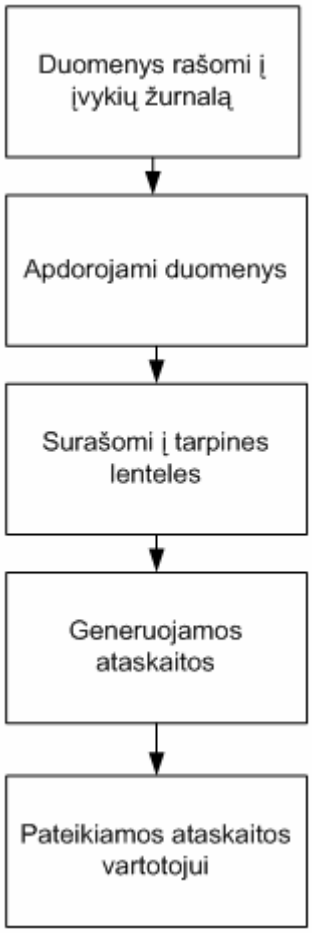
2.6.3 Ataskaitų generavimo būdai

Ataskaitų generavimą galima įsivaizduoti, kaip procesą, kurio metu iš visų duomenų yra susisteminama informacija ir pateikiama galutiniam vartotojui.



9 pav. Paprasčiausias ataskaitų generavimo mechanizmas

Galimi du būdai kaip visa turima informacija yra apdorojama. Pirmas variantas, kai kiekvieną kart peržiūrint ataskaitas jos yra generuojamos iš pradžių (9 pav.). Kitas variantas, kai duomenys tam tikrais laiko tarpais yra dalinai apdorojami ir surašomi į tarpines lenteles, o tik po to iš tarpinių duomenų generuojamos galutinės ataskaitos (10 pav.). Naudojamu tarpiniu duomenų bazės lentelių struktūra priklauso nuo generuojamos ataskaitos tipo, kartais skirtingoms ataskaitoms gali reikti skirtingos tarpinių lentelių struktūros. Esant dideliam duomenų srautui naudojimasis tarpinėmis lentelėmis užtikrina greitesnį sistemos darbą.



10 pav. Duomenų apdorojimas naudojant tarpines lenteles

Jei duomenys atnaujinami tam tikrais laiko tarpais iškyla problema, kad ne visada bus naujausi duomenys. Galima saugoti tik dalinai apdorotus duomenis, ištrinant jau apdorotus iš bendrų sisteminių įvykių žurnalų, tuo atveju sutaupoma saugojimo vietos, tačiau vėliau prireikus, generuoti kitokio tipo ataskaitas, gali reikti neapdorotų duomenų.

2.6.4 Ataskaitų atnaujinimo dažnumas ir resursai

Iš didelio informacijos kiekio generuoti ataskaitas kiekvieną kart jas peržiūrint gali pareikalauti didelių sistemos resursų. Todėl patartina informaciją apdoroti tam tikrais laiko tarpais apdorojant tam tikro laikotarpio informaciją. Tam reikalui būtina susikurti tarpines lenteles duomenų bazėje, į kurias būtų įrašomi duomenys apibendrinantys tam tikrą laikotarpį. Tačiau būna atvejų kai reikalinga papildyta ataskaita apie ką tik padarytus vartotojo veiksmus. Tada reikia kiekvieną kartą generuoti ataskaitas iš visų duomenų. Yra trys kintamieji kurie turėtų padėtį priimti sprendimą dėl ataskaitų atnaujinimo dažnio tai – poreikis tokioms ataskaitoms, turimų skaičiavimo resursų pajėgumui, bei portalo lankytojų skaičius. Galimas informacijos apdorojimas ir po kiekvieno vartotojo veiksmo, tačiau tada pats portalo veikimas gali pareikalauti nemažai sistemos resursų. Jei yra paskiriamas atskiras

analizės serveris, tada galima panaudoti jį informacijos apibendrinimui realiame laike, ir tada gauti ataskaitas pagal šviežiausius duomenis. Galima išvesti tiesinę priklausomybę, tarp sunaudojamų sistemos resursų (procesoriaus, atminties) ir ataskaitų atnaujinimo dažnumo.

2.7 Portalų analizės tipai

Šiuo metu analizės sistemos siūlo labai įvairų spektrą, ataskaitų. Kai kurios yra unikalios tik kai kuriems sistemų gamintojams. Tačiau visas ataskaitas galima skirstyti į kelias grupes, vienos yra susijusios su e-komercija, kitos su vartotojo sąsaja. Kadangi visų analizių neįmanoma aprėpti, todėl buvo nuspręsta šiame darbe apžvelgti tik kelių tipų analizę.

2.7.1 Lankomumo analizė

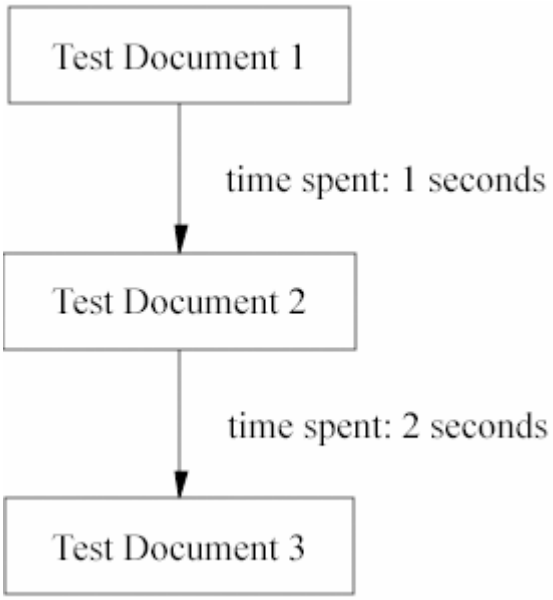
Lankomumo analizė yra apsilankymų per laikotarpį analizė. Ši analizė gali būti atliekama pagal labai daug kriterijų. Vienas iš kriterijų yra lankytojų skirstymas pagal demografinę ar kitokią informaciją, pagal kurią galima grupuoti lankytojus. Kitas šios analizės kriterijus yra laiko tarpas per kurį įvyko apsilankymai. Keičiant šios statistinės informacijos pasiskirstymą ir yra sudaromos ataskaitos.

Pagrindinė problema iškylanti sudarant šias ataskaitas yra iš registruotų įvykių atrinkti ir susisteminti informaciją. Vartotojų naršymo svetainėje metu, kai kurie puslapiai gali būti peržiūrėti keletą kartų, todėl bendroje statistikoje tai gali iškreipti informaciją. Todėl reikia atskirti sąvokas puslapio peržiūra ir apsilankymas. Puslapio peržiūra tai vieno puslapio informacijos peržiūra. Apsilankymas tai lankytojo veiksmai, kurio metu jis peržiūri portalo puslapius, jo apsilankymas baigiasi tada kai vartotojas baigia portalo naršymą.

Suregistruoti įvykiai yra nestruktūrizuoti duomenys, todėl prieš generuojant ataskaitas reikalinga atlikti jų struktūrizaciją, pagal reikalingus kriterijus, vartotojo demografine informacija, laiko intervalus, puslapius ir panašiai. Kriterijai yra pasirenkami pagal poreikius.

2.7.2 Navigacijos analizė

Šios analizės metu yra analizuojama vartotojo elgsena portale. Vartotojo naršymo ypatumai. Šios analizės metu yra svarbus puslapių peržiūros eiliškumas, laiko tarpai kiek puslapiai yra peržiūrimi. Navigacijos analizės metu yra bandoma sudaryt grafus atvaizduojančius puslapių peržiūrą. Individualus puslapio peržiūros grafas pavaizduotas 11 pav. Individualus navigacijos grafas



11 pav. Individualus navigacijos grafas

Sudarant navigacijos virsnės yra puslapiai, o briaunomis sujungti galimi navigacijos keliai. Briaunų vertė yra vidutinis laiko tarpas tarp vieno puslapio peržiūros ir kito. Pagrindinis uždavinys realizuojant šią navigacijos analizę yra duomenų transformavimas iš nestruktūrizuotų duomenų į struktūras palengvinančias tokių ataskaitų generavimą.

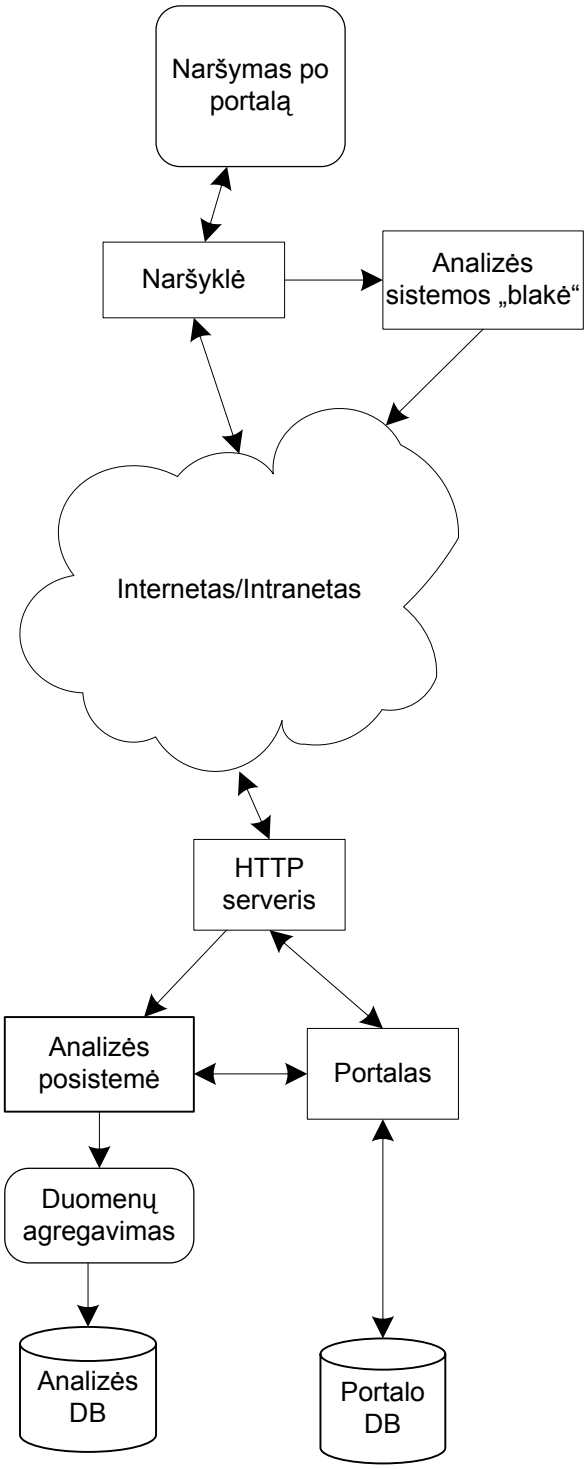
3. Portalo duomenų apdorojimo ir analizės sistemos teorinis modelis

Sistemos projektavimo tikslas yra – sistema galinti rinkti, analizuoti informaciją ir pateikti ataskaitas apie portalą. Tokios sistemos pagrindiniai reikalavimai yra:

- Duomenų apdorojimas ir analizė realiame laike.
- Portalo ir analizės duomenų integracija.
- Duomenų saugojimo ir kaupimo problemos sprendimas.

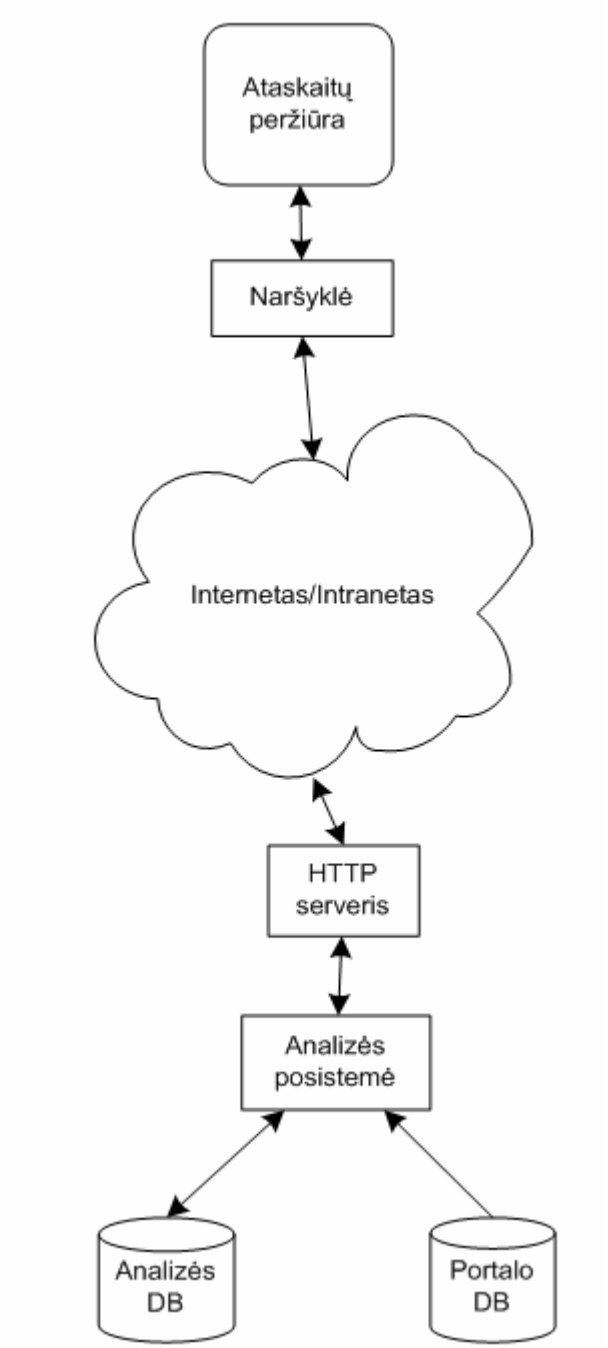
3.1 Sistemos architektūra

Išanalizavus portalo duomenų apdorojimo ir analizės sistemų kūrimo metodus ir technologijas prieita išvadų, kad geriausiai tiktų hibridinis modelis. Jo koncepcija pavaizduota 12 pav. Duomenys renkami naudojant kelis būdus. Yra panaudojamos JavaScript „blakės“ įterptos į kiekvieną portalo puslapį, bei sistema registruojanti įvykius portalo viduje. Analizės sistema duomenis gauna vienu formatu, juos apdoroja ir į duomenų bazę, rašo į formatą labiau tinkamą ataskaitų generavimui. Tokiu būdu ataskaitų generavimas yra iš pačių naujausių duomenų.



12 pav. Sistemos modelis, renkant duomenis

Peržiūrint ataskaitas (žr. 13 pav.) imami jau apdoroti duomenys iš analizės duomenų bazės, jei reikia imami duomenys ir iš portalo duomenų bazės.



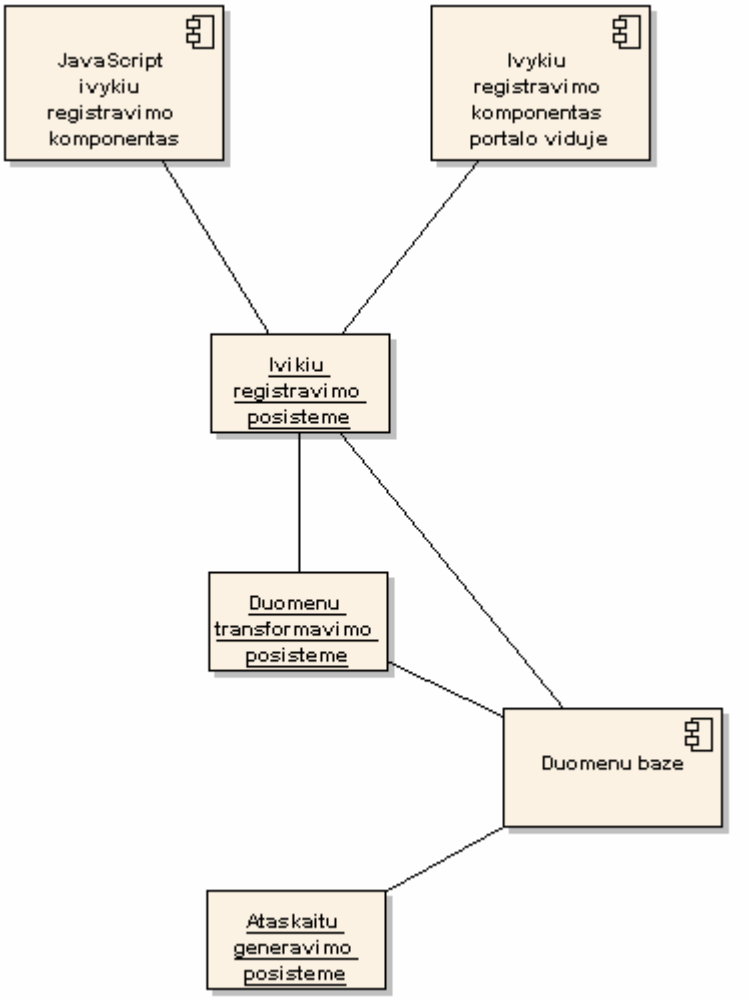
13 pav. Sistemos modelis, analizuojant duomenis

3.2 Duomenų struktūros

3.2.1 Duomenų srautai

Duomenų srautas registruojant įvykius yra tik viena kryptimi. Duomenys registruojami iš duomenų šaltinių, šiuo atveju iš JavaScript blakių, kurios yra įvykdomos portalo viduje ir iš sistemos, kuri registruoja įvykius portalo viduje. Abu pirminiai duomenų srautai yra nepriklausomi vienas nuo kito, tačiau duomenų transformavimo metu jie yra apjungiami. 14 pav. Sistemos komponentinis modelis pavaizduotas sistemos komponentinis modelis. Diegimo metu į veikiantį portalą yra įterpiami du komponentai: sukuriantis „blakę“ ir

registruojantis vidinius portalo duomenis. Įvykių registravimo posistemė, surinktus duomenis persiunčia duomenų transformavimo posistemei, kuri transformuoja duomenis ir rašo juos į duomenų bazę. Ataskaitų generavimo posistemė skaito iš duomenų bazės jau transformuotus duomenis, ir generuoja ataskaitas.



14 pav. Sistemos komponentinis modelis

Pagrindiniai šios sistemos komponentai:

- Įvykių registravimo posistemė

  Pagrindinė funkcija užtikrinti duomenų rinkimą vartotojui naršant portale. Šis komponentas surinktą informaciją, naudodamas duomenų bazę arba tiesiogiai, perduoda informaciją duomenų transformavimo posistemei.

- Duomenų transformavimo posistemė

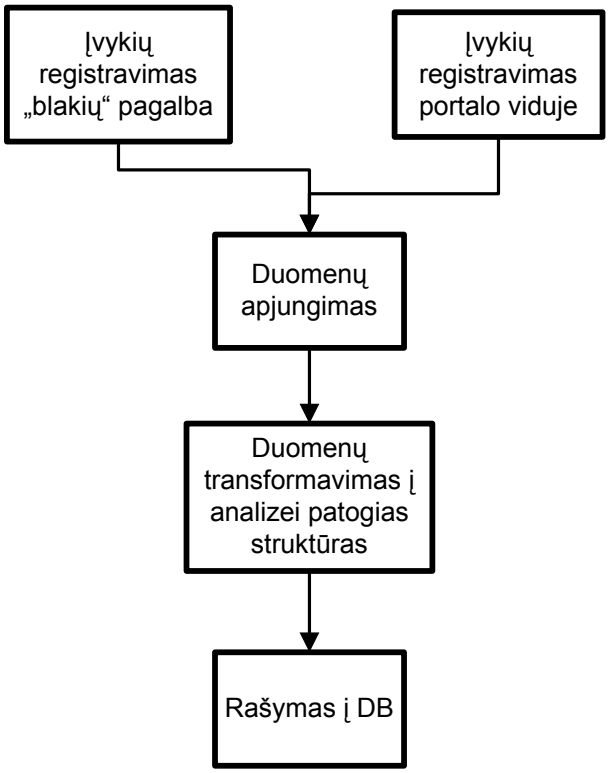
  Pagrindinė šios posistemės funkcija transformuoti surinktus duomenis į duomenų formatą patogesnį ataskaitų generavimai. Ši posistemė gali veikti nepriklausomai nuo įvykių registravimo posistemės. Transformuoti duomenys įrašomi į duomenų bazę.

- Ataskaitų generavimo posistemė

Pagrindinė funkcija sugeneruoti grafikų ar lentelių pavidalu ataskaitas apie portalą. Ši posistemė naudoja jau transformuotus duomenis, jei reikia gali imti ir netransformuotus. Taip pat naudoja ir portalo duomenis.

3.2.2 Įvykių registravimo posistemė

Kuriamoje sistemoje yra keletas duomenų srautų. Pirminiai duomenys yra renkami iš kelių duomenų šaltinių. Pirminiai duomenys savo formatu yra nuoseklus vartotojų veiksmų registravimas. Vartotojo veiksmai yra tokie, kaip nuorodų paspaudimas, duomenų įvedimas į formas. Kiekvieno vartotojo veiksmo metu yra sugeneruojamas įvykis kuris yra apdorojimas analizės posistemės. Analizės posistemė transformuoja duomenis į formatą leidžiantį greičiau generuoti ataskaitas.



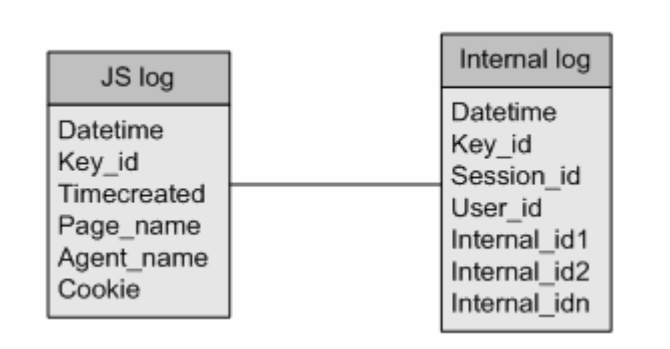
15 pav. Įvykių registravimo eiga

15 pav. Įvykių registravimo eiga pavaizduota įvykių registravimo eiga. Įvykiai registruoti abiejuose duomenų šaltiniuose yra apjungiami, vėliau struktūrizuojami ir įrašomi į duomenų bazę.

3.2.3 Duomenų transformavimo posistemė

Duomenų transformavimas yra pradinių duomenų formos pakeitimas į formą patogesnę atlikti galutinės analizės ataskaitų generavimui. Iš apdorotų duomenų yra daug paprasčiau ir

greičiau galima generuoti ataskaitas. Pradinių duomenų struktūrą apibrėžia duomenų šaltiniai. Kadangi projektuojamoje sistemoje yra naudojami du duomenų šaltiniai tai iš abiejų surenkami duomenys, kurie yra svarbiausi tolimesnei analizei.



16 pav. Pradinio formato duomenų struktūra

15 pav. pavaizduota pradinė duomenų struktūra kuria yra renkami duomenys. **JS log** duomenys gaunami iš JavaScript „blakės“ (3 lentelė), **Internal log** duomenys gaunami iš portale integruoto duomenų surinkimo mechanizmo (4 lentelė). *Key_id* laukas suriša duomenis gaunamus iš dviejų duomenų rinkimo šaltinių.

3 lentelė JS log duomenų struktūros paaiškinimas

Datetime	Laikas kada įrašas įterptas į db
Key_id	Identifikacinis raktas surišantis su **Internal log**
Timecreated	Laikas kada buvo sukurta blakė serveryje
Page name	Puslapio pavadinimas
Browser	Lankytojo naudojamos naršyklės pavadinimas
Cookie	Sausainėlis įterptas vartotojui

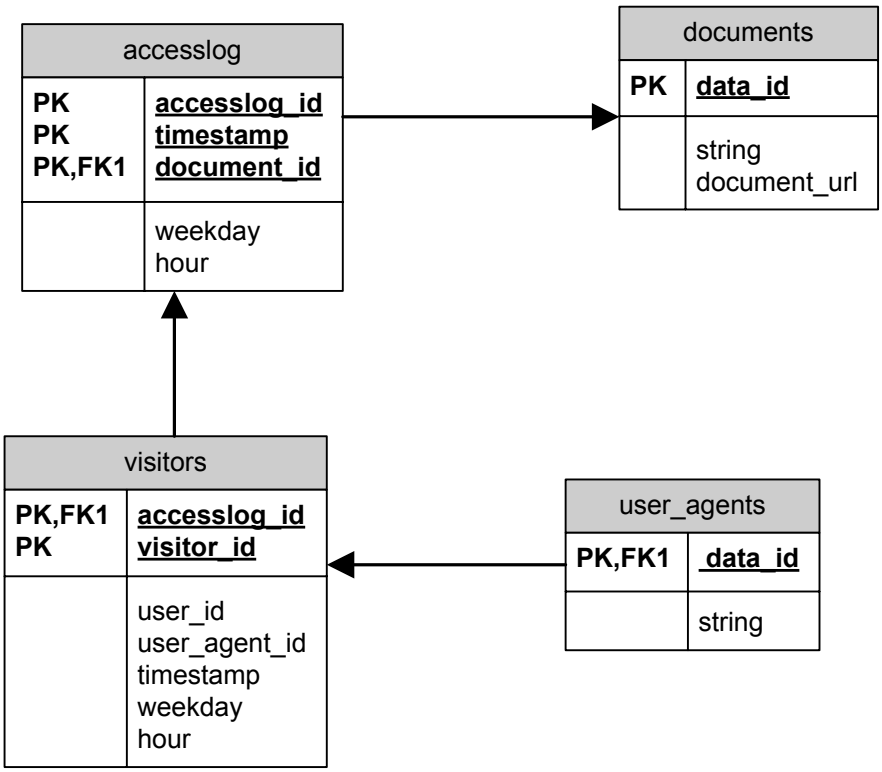
4 lentelė Internal log duomenų struktūros paaiškinimas

Datetime	Laikas kada įrašas įterptas įterptas į db
Key_id	Identifikacinis raktas surišantis su **JS log**
Session_id	Laikas kada buvo
User_id	Vartotojo ID naudojamas portalo vidinėje sistemoje
Inernal_id1	Portalo vidinio resurso ID
Inernal_id1	Portalo vidinio resurso ID
Inernal_id1	Portalo vidinio resurso ID

Tokio duomenų rinkimo privalumas tas, kad naudojant JavaScript „blakes“ galima surinkti duomenis, kuriuos galima gauti tik vykdant kodą lankytojo naršyklėje. O vykdant kodą serveryje, galima gauti duomenis susijusius su portalo vidiniais resursais, kurie yra svarbūs duomenų integracijai.

3.2.3.1 Lankomumo analizės duomenų transformacija

Struktūrizuojant lankomumo duomenis lankomumo analizei reikia įvertinti kriterijus pagal, kuriuos gali būti klasifikuojami duomenys. Šiuo atveju duomenys klasifikuojami pagal *Page name, Browser*. Lankytojai identifikuojami pagal kelis parametrus *user_id* ir *Cookie*. Struktūrizuojant duomenis, duomenų bazės abstrakčiame lygyje kiekvienam kriterijui yra sukuriama atskira lentelė. Duomenų struktūra saugoti transformuotiems duomenims pavaizduota 17 pav. Duomenų struktūra saugoti transoformuotiems duomenims



17 pav. Duomenų struktūra saugoti transoformuotiems duomenims

5 lentelė accesslog struktūros paaiškinimas

accesslog_id	Įvykio unikalus id
timestamp	Laikas, kada įvykis buvo registruotas
document_id	Puslapis, kuris buvo peržiūrėtas
weekday	Savaitės diena
Hour	Valanda

6 lentelė documents struktūros paaiškinimas

data_id	Puslapio unikalus id
String	Puslapio pavadinimas
document_url	Puslapio URL

7 lentelė visitors struktūros paaiškinimas

accesslog_id	Įvykio unikalus id
visitor_id	Lankytojo unikalus id
user_id	Puslapis, kuris buvo peržiūrėtas
user_agent_id	Naršyklės id
weekday	Savaitės diena
Hour	Valanda

8 lentelė user_agents struktūros paaiškinimas

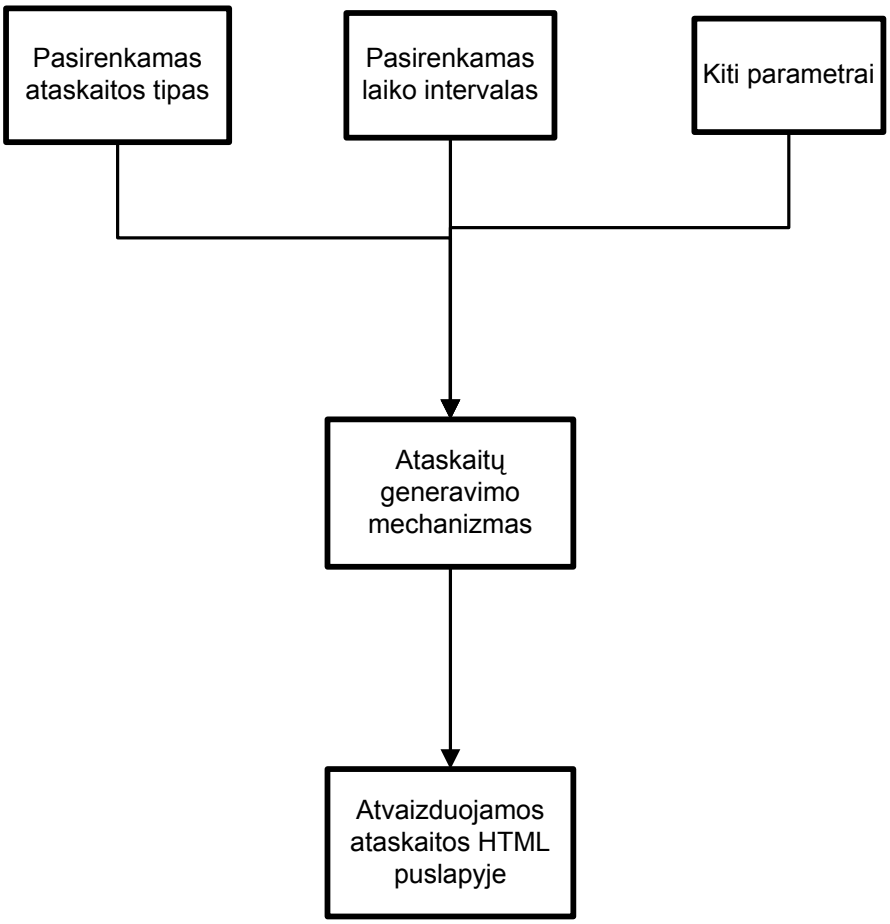
data_id	Unikalus naršylės id
String	Puslapio pavadinimas
document_url	Puslapio URL

Tokios transformuojamos duomenų struktūros patogumas grindžiamas tuo, kad generuojant ataskaitas nereikia iš bendros masės informacijos atrinkinėti pagal tam tikrą kriterijų. Pvz. Visi puslapiai registruojami atskiroje lentelėje, todėl generuojant ataskaitą, kuris puslapis buvo peržiūrėtas daugiausia kartų, tereikia suskaičiuoti įrašus iš **accesslog** lentelės grupuojant pagal *document_id*.

3.2.4 Ataskaitų generavimo posistemė

Ataskaitų generavimo proceso metu, iš apdorotų statistikos duomenų yra sukuriamos ataskaitos lentelių, diagramų ar grafikų pavidalu. Dažniausiai naudojamos diagramos, kad aiškiau būtų matomas pasiskirstymas. Lankomumo ataskaitoje yra svarbu parodyti, koks buvo bendras apsilankiusių lankytojų skaičius, per tam tikrą laikotarpį. Jei tas laikotarpis yra diena, tai diagrama turi būti suskirstyta valandomis, jei pasirinktas laikotarpis savaitė – savaitės dienomis. Šioje sistemoje galima generuoti tokias lankomumo ataskaitas.

- Labiausiai peržiūrimi puslapiai.
- Vartotojų apsilankymų skaičius tam tikrame puslapyje.
- Vartotojų apsilankymų skaičius tam tikrame puslapyje, per tam tikrą laikotarpį.
- Populiariausios naršyklės.



18 pav. Ataskaitų generavimo principinė schema

Prieš generuojant ataskaitas yra pasirenkamas ataskaitos tipas, priklausomai nuo ataskaitos tipo gali būti pasirinktas laiko tarpas ar kiti detalizuojantys parametrai. Tada kreipiamasi į ataskaitų generavimo mechanizmą, kuri iš apdorotų duomenų ir portalo duomenų sugeneruoja ataskaitas HTML formatu.

3.3 Sistemos modelio apibendrinimas

Sistemos privalumai:

- Pasiūlytu duomenų surikimo modeliu galima surinkti visus analizei reikalingus duomenis.
- Naudojant hibridinę duomenų rinkimo sistemą, galima išmatuoti laiką, per kiek laiko, puslapis pasiekė vartotoją.
- Renkami duomenys iškart yra apdorojami todėl, nėra kaupiami visi įrašai, kas sumažina duomenų bazės dydį.

Sistemos trūkumai:

- Kadangi saugomi tik apdoroti duomenys, negalima lengvai pakeisti generuojamų ataskaitų.

- Sistema priklausoma nuo programavimo kalbos kuria parašytas portalas.

4. Tiriamos sistemos struktūros ir realizacijos tyrimas

Projektiniai daliai atlikti yra naudojamas KTU kompiuterių katedros modulių informacinė sistema, kuri buvo autoriaus sukurta bakalauro darbo metu. KTU kompiuterių katedros modulių portalas yra jau padaryta veikianti sistema. Projektuojama portalo duomenų apdorojimo ir analizės sistema yra atskirai veikianti sistema. Tarp jų komunikacijai į analizuojamą portalą yra įterpiama duomenų surinkimo „blakė“. Pagrindinis uždavinys šiai sistemai sugebėti sekti atskirų portalo vidinių resursų naudojimo statistiką. Vidiniai portalo resursai yra moduliai, vartotojai, paskaitos ir panašiai.

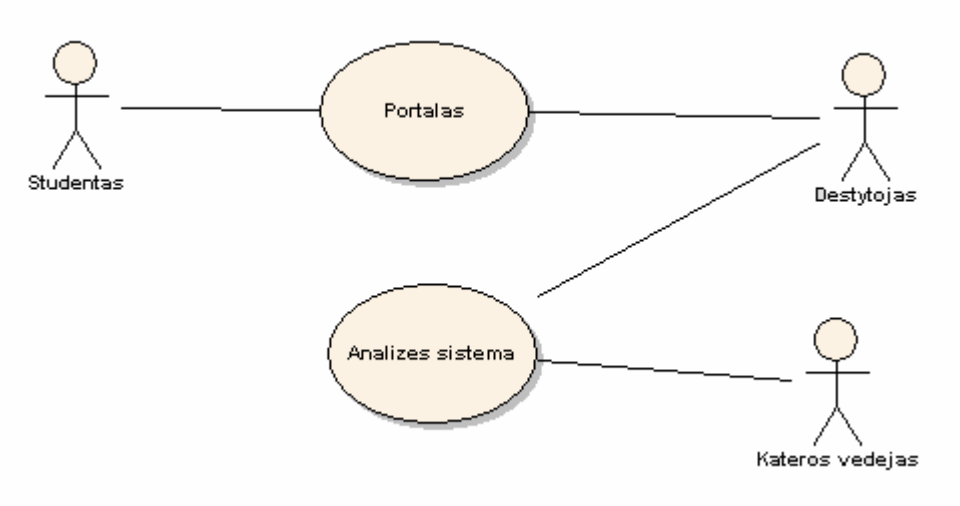
4.1 Situacijos analizė

Kompiuterių katedra yra Kauno Technologijos Universiteto Informatikos fakulteto padalinys. Kaip ir bet kuri kita organizacija, kompiuterių katedra siekia didinti darbo kokybę ir našumą. Tai padaryti yra daug būdų. Vienas iš jų, kuris mums aktualiausias yra veiklos kokybės gerinimas. Pagrindiniai kompiuterių katedros veiklos dalyviai yra dėstytojai ir studentai. Pagrindinė katedros veikla yra žinių perdavimas studentams, bei dėstytojų profesinis tobulėjimas.

Veikiant informacijos mainam tarp dėstytojų ir studentų yra aktualu žinoti ar portalas yra naudojamas pagal paskirtį, t.y. studentai skaito portale padėtą informaciją, dėstytojai atnaujina patalpintą medžiagą.

4.1.1 Veiklos sąveikų modelis

Pagrindiniai organizacijos veiklos dalyviai yra studentas, dėstytojas ir katedros vedėjas. Studentai negali peržiūrėti jokios statistikos. Dėstytojai gali peržiūrėti jų dėstomų modulių studentų statistiką. Katedros vedėjas gali peržiūrėti viską.



19 pav. Veiklos sąveikų modelis

4.1.2 Portale pateikiama informacija

Portale yra pateikiama informacija, kurios naudojimą reiktų stebėti:

- Moduliai
- Skelbimai
- Paskaitų temos
- Laboratoriniai darbai
- Tvarkaraštis
- Įvertinimai

Dėstytojui yra svarbu žinoti ar studentai lankosi portale ir naudojasi pateikiama medžiaga. Katedros vedėjui svarbi informacija kaip bendrai naudojamas portalas tiek studentų, tiek dėstytojų.

4.1.3 Reikalingos ataskaitos portalo analizei

Šiam portalui analizuoti atsižvelgiant į pateikiamą informaciją reikalingos tokios ataskaitos:

- Bendras vartotojų apsilankymas portale.
- Vartotojų apsilankymas tam tikruose moduliuose.
- Vartotojų apsilankymas tam tikruose modulio dalyse (skelbimai, paskaitų temos, laboratoriniai darbai, tvarkaraštis, įvertinimai).
- Labiausiai naudojamos naršyklės, ekrano rezoliucijos ir pan.

Taip pat visose ataskaitose galima rūšiuoti informaciją pagal vartotojų informaciją (dėstytojas, studentas, grupė, kursas ir pan.). Kai kuriuose ataskaitose galima nurodyti laiko

intervalą, tai atvaizduojant lankomumą vizualiai matosi kuriuo paros metu, ar savaitės dieną lankomumas buvo didžiausias.

4.1.4 Resursų ir vartotojų identifikavimas

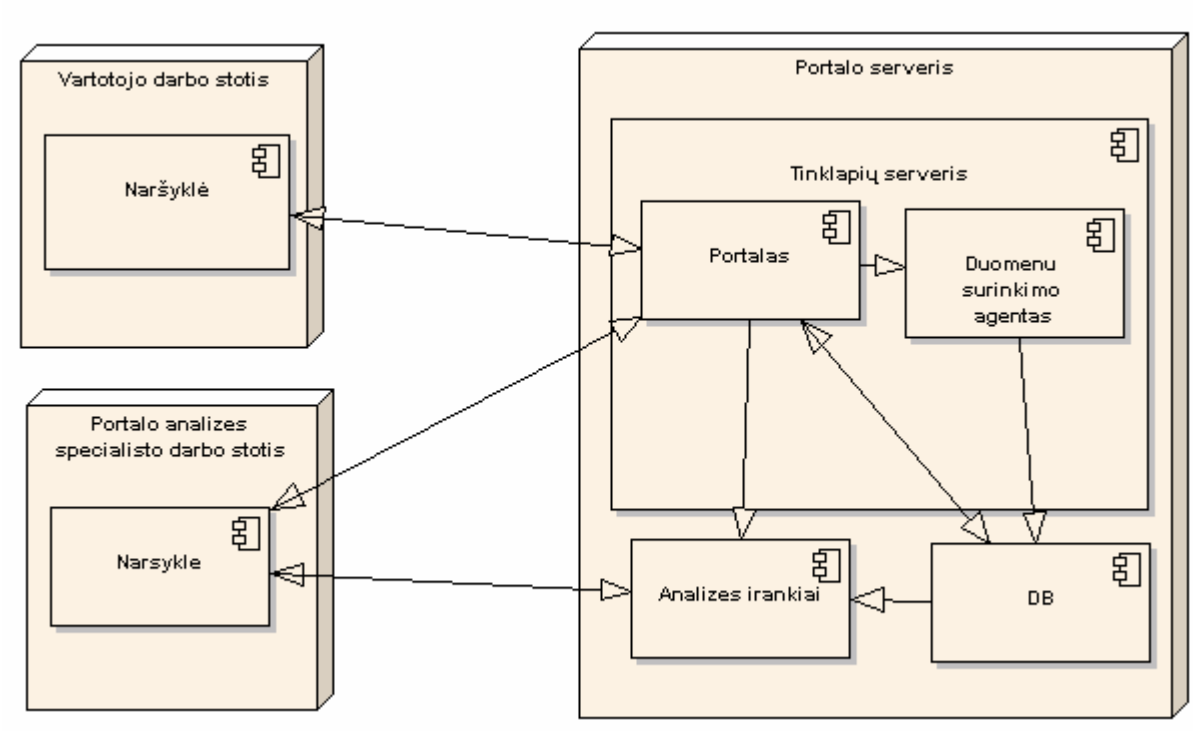
Visi lankytojai, kurie naudojasi šiuo portalu, turi prisijungti, naudodamiesi savo vartotojo vardu ir slaptažodžiu. Tai palengvina lankytojų identifikaciją, kadangi nereikia atskirai identifikuoti unikalių lankytojų, o tai galima padaryti naudojantis portalo sistemoje esančiu unikaliu vartojo identifikatoriumi.

Atskiri sistemos resursai (moduliai, paskaitų temos ir pan.) taip pat identifikuojami naudojantis vidiniais sistemos identifikatoriais.

4.2 Sistemos realizacija

Portalo sistema suprogramuota Java programavimo kalba. Naudojama duomenų bazė – MySQL. Todėl analizės sistema realizuojama naudojantis tomis pačiomis priemonėmis.

4.2.1 Komponentų diagrama

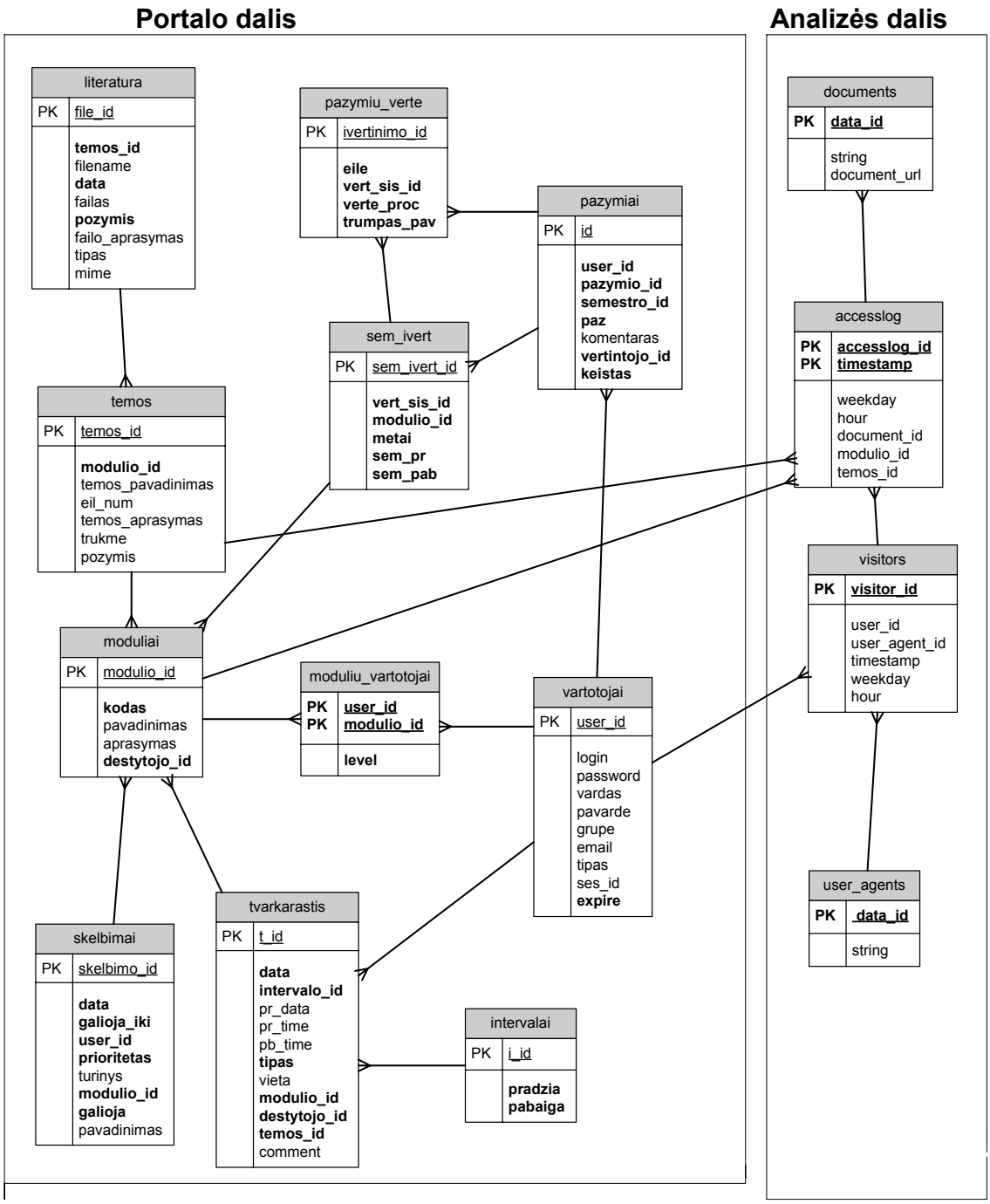


20 pav. Sistemos komponentų diagrama

Sistemos komponentų diagrama pavaizduota (20 pav.). Analizės sistema veikia tinklapių serveryje, kurio pagalba yra surenkami duomenys, bei analizuojami. Ataskaitoms peržiūrėti yra naudojama naršyklė

4.2.2 Duomenų bazės fizinis modelis

Portalo ir analizės sistemos duomenų bazės struktūra pavaizduota 21 pav.



21 pav. Duomenų bazės struktūra

Diegiant sistemą, atskiroje duomenų bazėje sukuriamos papildomos lentelės statistikos duomenims. Analizės sistema taip pat naudoja ir portalo duomenis. Portalo duomenys panaudojami vartotojų identifikavimui, informacijos apie vartotojus ar kitus portalo resursus ataskaitų detalizavimui.

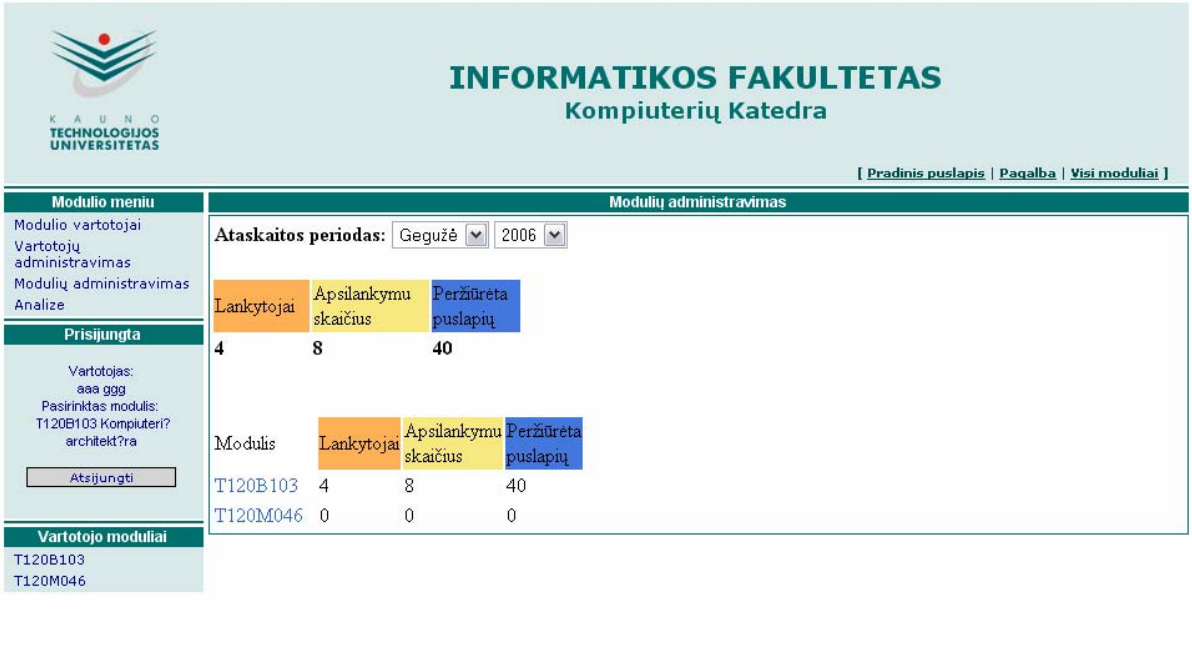
4.3 Sistemos bandomasis veikimas

Duomenų rinkimui į pagrindinį failą *main.jsp* įterpiama kodo dalis:

```
surinkimas.wirte_log(request,user_id,modulio_id,
 temos_id);
```

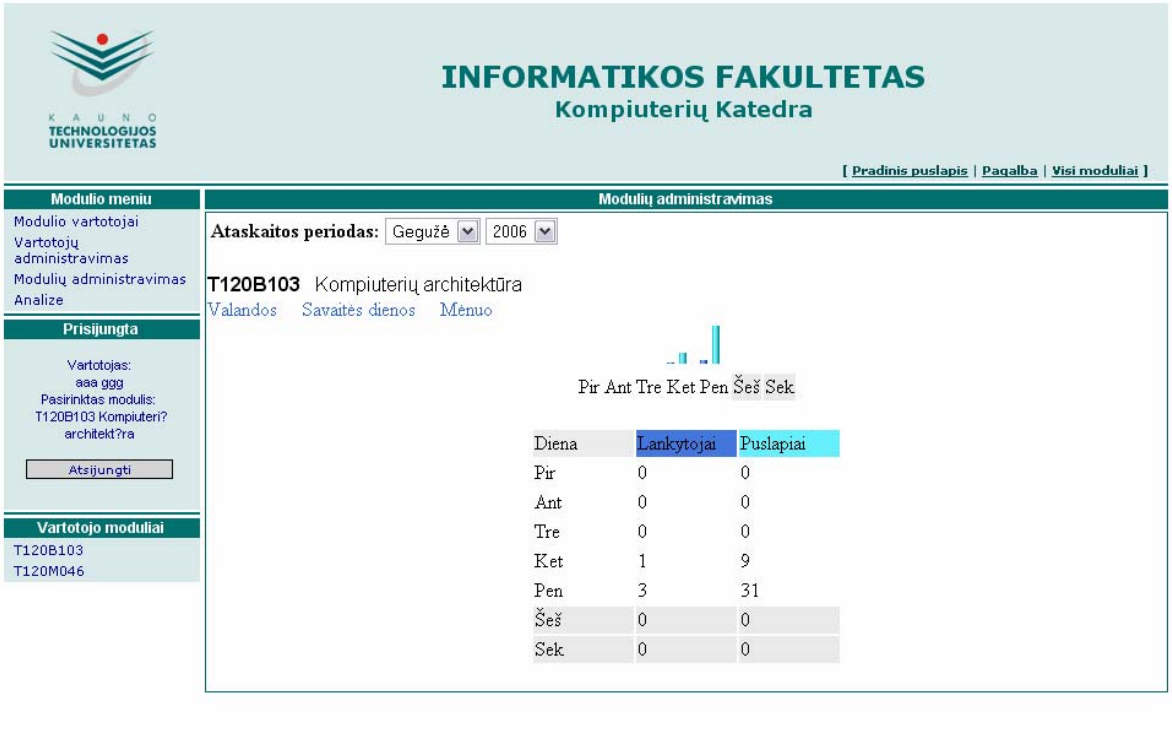
Įterpta kodo dalis renka duomenis iš užklausos, bei registruoja vidinių portalo resursų identifikatorius. Tai pat ši kodo dalis sugeneruoja ir JavaScript kodo dalį, papildomos informacijos surinkimui.

Portalo analizės sistema paleidžiama prisijungus administratoriaus slaptažodžiu prie sistemos. Atsiranda papildomas meniu punktas „Analizė“. Paspaudus šį meniu punktą matomas vaizdas 22 pav.



22 pav. Bendri statistikos duomenys

Šiame paveiksle matoma bendras lankytojų skaičius apsilankiusių per pasirinktą periodą. Taip pat detalizuota informacija apie konkrečius modulius. Pasirinkus modulį galima sužinoti smulkesnę informaciją apie jį. 23 pav. pavaizduota detalesnė lankomumo informacija sugrupuota pagal savaitės dieną. Informaciją taip pat galima grupuoti pagal valandas, pagal mėnesio dienas.



23 pav. Detalesnė modulio informacija

4.4 Bandomosios sistemos įvertinimas

Suprojektuota sistema buvo bandoma tame pačiame serveryje, kuriame buvo ir kuriama. Sistema veikė stabiliai. Buvo stebima ar sistema atitinka reikalavimus keliamus tokioms sistemoms:

- Sistema tiksliai nustato lankytoju skaičių.
- Ataskaitos generuojamos iš naujausių duomenų.
- Sistema nekaupia visų registruotų įvykių, tik apibendrintus duomenis.
- Sistema gali analizuoti portalo vidinių resursų lankomumą.

5. Išvados

- Išanalizavus galimus duomenų surinkimo šaltinius, kad bendru atveju patogiausia naudoti kelis duomenų šaltinius iškart. Naudingiausia vienu metu naudoti vieną šaltinį kurio kodas yra vykdomas kliento naršyklėje, ir kitą kurio kodas yra vykdomas serveryje. Tokiu būdu galima surinkti didžiausią kiekį duomenų.
- Šiame darbe apžvelgta kokiais kriterijais remiantis yra analizuojami portalai. Nustatyta, kad svarbiausi portalo analizės elementai yra lankytojai ir puslapiai. Apibrėžtos problemos, kurios gali iškilti bandant analizuoti šiuos kriterijus, bei pasiūlyti šių problemų sprendimai.
- Pasiūlyta konceptualus duomenų apdorojimo ir analizės modelis, naudojantis hibridinį duomenų surinkimo šaltinį.
- Išanalizuotas, projektiniai daliai pasirinktas, KTU kompiuterių katedros portalo studijų modulių sistema. Nustatytas poreikis kokių gali reikti ataskaitų analizuojant šį portalą.
- Realizuota sistema buvo patikrinta ar atitinka reikalavimus keliamus portalų duomenų apdorojimo ir analizės sistemoms.
- Sistemą reiktų labiau tobulint stebint ją jos eksploatacijos eigoje.

6. Literatūra

1. Eric T. Peterson. Web Site Measurement Hacks. Boston: O'Reilly, 2005.

2. Jim Stern. Web Metrics-Proven Methods for Measuring Web Site Success. Denver: Willey Publishing, Inc., 2002.

3. Eric T. Peterson. Web Analytics Demystified. Celio Group Media, 2004.

4. Web analytics move across the enterprise. 2001 [Žiūrėta 2005-12-14]. Prieiga per internetą: http://www.infoworld.com/articles/fe/xml/01/08/20/010820feedge.html.

5. Mark Sweiger, Mark Madsen, Jimmy Langston, and Howard Lombard. Clickstream Data Warehousing, Willey Publishing, Inc., 2002.

6. Improve Your Communication To Improve Web Conversion. Prieiga per internetą: http://www.conversionchronicles.com

7. Brian Clifton. Web Traffic Data Sources & Vendor Comparison. Prieiga per internetą: http://www.omegadm.co.uk/web-analytics-methods.pdf

8. Jakob Neilsen. Web Usability: The practice of Simplicity. Indianapolis: New Riders Publishing, 2000

9. Analog:WWW log file analysis. Prieiga per internetą: http://www.analog.cx/

10. AWStats – Free log file analyzer for advanced statistics. Prieiga per internetą: http://awstats.sourceforge.net/

11. PhpOpenTracker: framework solution for the analysis of website traffic and visitor anglysis. Prieiga per internetą: http://www.phpopentracker.de/en/

12. Zoomstats – Traffic Analyzing to its best. Prieiga per internetą: http://zoomstats.sourceforge.net/

13. ClickTracks Web Analytics for PPC, SEO and ROI stats. Prieiga per internetą: http://www.clicktracks.com

14. Top100 - interneto sistema, pristatanti tinklalapių lankomumo statistiką bei jos analizę. Priega per internetą: http://top100.penki.lt/

7. Terminų ir santrumpų žodynas

9 lentelė Terminai

Lietuviškai	Angliškai	Paaiškinimas
Modelis	Model	Modelis yra sistemos dalies (funkcijos, struktūros ir/arba elgesio) atvaizdavimas kokioje nors sintaksiškai formalioje kalboje
Serveris	Server	Sistemos aparatūrinė dalis, aptarnaujanti sistemos programinius modulius.
Portalas	Portal	Interneto svetainė, pasižyminti didele informacijos bei papildomų paslaugų gausa, kuri gali tarnauti kaip išeities taškas į kitus interneto resursus

10 lentelė Santrunpos

Santrumpa	Pilnas pavadinimas	Paaiškinimas
CLF	Common Log Format	Sisteminių įvykių žurnalo failo formatas
URL	Uniform Resource Locators	Adresas iki internetinio resurso
URI	Uniform Resource Indentifier	Internetinio resurso identifikatorius
LAN	Local Area Network	Tinklo standartas
HTML	HyperText Markup Language	Publikavimo kalba internetui
HTTP	HyperText Transfer Protocol	Internetinis protokolas