Vadyba
Journal of Management
2009, Vol. 14, No. 1
ISSN 1648-7974

USING DATA WAREHOUSE AND AGENT TECHNOLOGIES TO PREPARE E.LEARNING DATA FOR DATA MINING

Jelena Mamčenko, Inga Tumasonienė
Vilniaus Gedimino technikos universitetas
el. paštas: [1]jelena@gama.vtu.lt, [2]tinga@gama.vtu.lt

Abstract
The aim of this work is the application of data mining technologies to e.learning groupware system's data. The software which is analyzed in this paper was developed by *IBM Lotus Notes/Domino* and characterizes non traditional database model. In such model data are stored as a single objects and this is a serious problem in a way of deploying data mining software. We suggest a new method in order to avoid problems such as document based model data collection, transformation, aggregation and filtering, which are considering on agent and data warehousing technologies.
Methods for registering and processing *Internet* data are fairly new. During the design of infrastructure for information technologies not enough attention has been paid to collecting data suitable for data mining analysis. All Internet data flows are caused by various devices. However, every device has native data formats and uses different algorithms, so the first stage of data analysis, data collection, becomes much more complicated.
In this paper we tried to present example of applying the data mining technology for e.learning data. Invoking created agent and created data collection it became possible to apply data mining technology methods to the e.learning system's document data. Considering peculiarity of documental database, methodology allowing to transform in the real time data from documental database into data warehouse, invoking agent technologies. Interpreted results are useful from the department point of view. It helped to change necessary document databases and in future create more personalized site and better site layout. In addition to that the analysis of servers' activity was performed and servers scheduling was changed to a conditional non - busy time. This let escape a big workload and increase the *gama* and *irma* servers' efficiency. The proposed agent deals with document collections and can work in automatic and manual regime.
KEYWORDS: e.learning, data mining, document based model, data warehouse, agent technologies

Introduction

Vilnius Gediminas Technical University e.learning system is based on the documental model and represents a group work organization system. Whereby using common databases and information dissemination tools, group activity processes are ensured.

Users of such organization system are enabled to access its services from any point of the world. This document management system is designed for numerous data base users. System's options and functions enable to extend its use to user management system, centralized process and information management tool and team/group work tool.

The aim of this system is to provide e.learning possibilities and enable students to access information from any place of the world and at any time, as well as, communicate with course tutors. System is distributed among several servers, and each of them has own function. Such systems store large quantities of information; once it is analyzed it could be useful for e-learning purposes. However, the management of such information requires powerful tools. Application of data mining technologies could be an appropriate solution in this situation (Anjewierden *et al*. 2007; Despotović *et al*. 2008; Hanna 2001).

E.learning system is being based on the distributed servers'system. It implies the necessity for servers to connect automatically and exchange the changes in the databases. Users access different servers, i.e. multi-protocol servers, data base servers. These servers (using *Lotus/Domino* technology) sustain complete spectrum of technologies and *Web* standards.

Due to the continuously growing number of students and the quantity of new databases, the number of connections to the data basis is continuously increasing. Students use a number of applications, access data basis through different protocols, which results in the server overload and increases the time necessary to process user queries.

In order to save precious servers' resources and continue providing consumers with high quality access to the database, it was decided to find a solution to maintain the system functionality and save servers' resources. To achieve this objective it was decided to apply data-mining technologies for e.learning users *log* files (Srikant, Yang 2001).

Gama is the main application server, which enables users to access the software located in different system. The *LDAP* directory's task is to provide a simplified path to the catalogue, as well as, the possibility to use *HTTP*. *Irma* server contains learning management system model and one more application server.

Server *kappa* is dedicated for official use, i.e. it contains the e.learning management system and the *Lotus Notes* client. For this purpose it is necessary to create a data warehouse in the *kappa* server and so called *agent*, which would allow to automate complex actions, i.e.

collect data from document database and in the same time transform and prepare it for data mining analysis, create data mining software server, which would allow to perform data mining (Fig. 1).

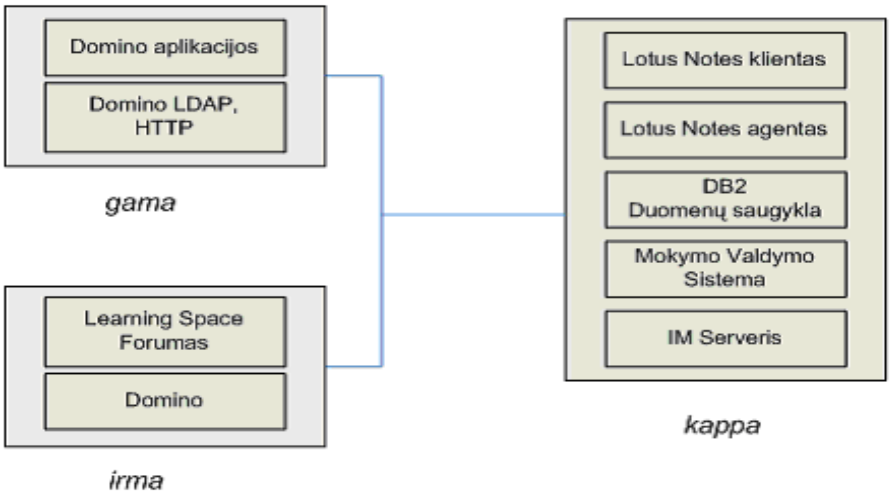

Figure 1. Interaction among distributed servers

Data mining technologies in document management system

VGTU e.learning system is created on the basis of document model and is a system dedicated to organize group work, where by shared use databases and information dissemination among a number of users the principles of group work processed are insured.

The main goal of this system is to confer learning courses on-line, it also allows to access necessary information from any place and at any time and communicate with course tutors.

Usually information is stored in servers (*Apache* and etc.) in text type. In our information system all information is being collected in the data base of each server (Fig. 2), its data format is not adapted to perform data mining.

By the help of the agent, information from each server it is being transferred to the created data warehouse, and during this process it is being transformed, cleaned from irrelevant data, stored and processed.

It is advisable to use agent technologies, which facilitate the implementation of the processes represented in the Figure 3.

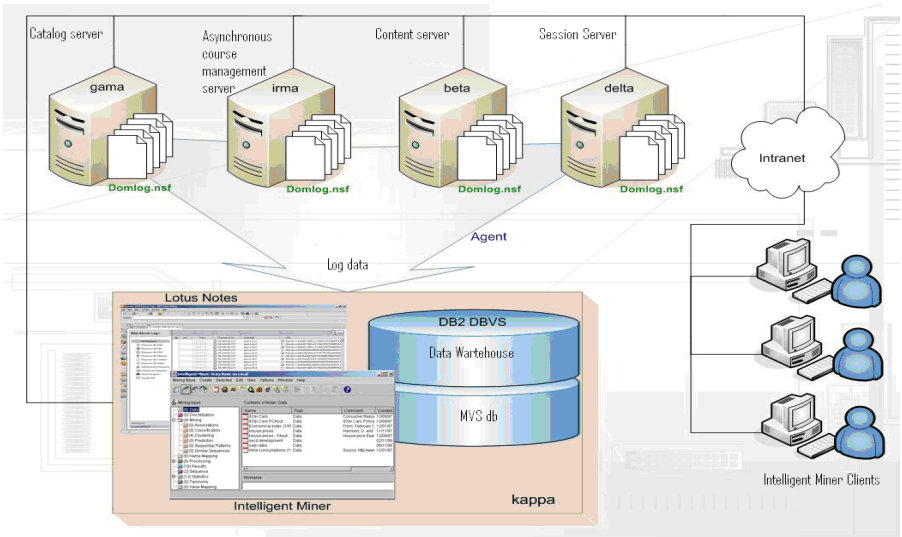

Figure 2. Usage data integration, storage and processing of the distance learning system

Data mining technology application for document databases needs new data transformation method, which could help to put all documents from document database into data warehouse. Existing software systems works only with traditional relational databases or with flat files. That's why we need to create a new transformation method for document databases data which have to be analyzed using data mining techniques. In this paper we

proposed new tool for document data transformation for data mining software using agent technology.

Historical business data is vital for strategic planning of activities and assessment of achieved results of separate departments.

To avoid problems discussed before and conserve historical data, separate data warehouses shall be created. All historical information shall be transferred to the warehouses, so server load would be smaller.

This problem was solved by using an *agent* that works with a collection of documents and inputs operational data in a data warehouse in real time.

Thus, the first task of the created *agent* (Fig. 3) is to select the database, which will release the *agent* and get all the documentary base documents to the document collection. The script of proposed agent was created using *LotusNotes/Domino* native language *Lotus Script* and the fragment is present in Figure 4.

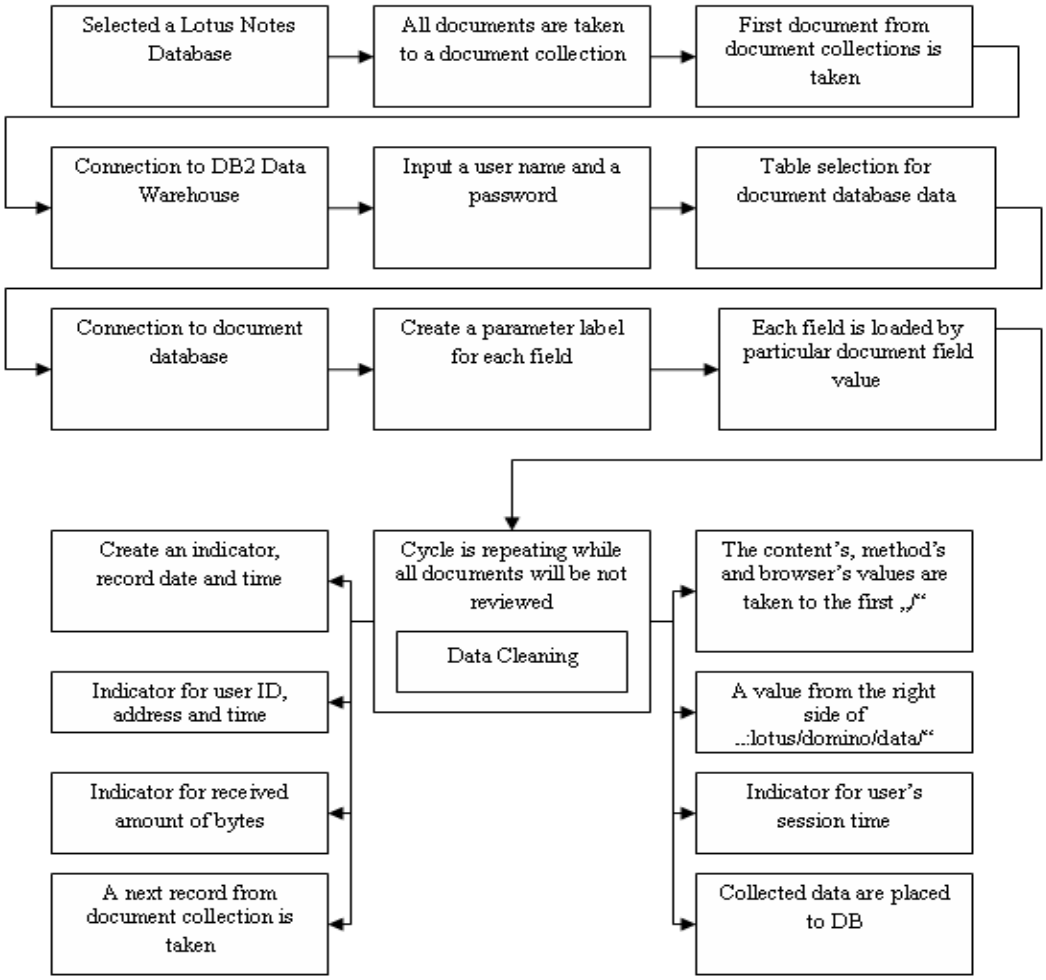

Figure 3. The algorithm of proposed agent

Then the document is taken from the collection and access to the DB2 data warehouse is being made. In the next step, each field is given specific column settings. After inclusion of the relevant parameters in the field, their values are inputted. From here the whole cycle is repeated till all the documents are processed. Data cleaning is performed at this stage.

Use of the agent ensures that data is being moved to the warehouse, where the operational data is securely stored, after data integration, filtering and structuring has been carried out. Currently, the data warehouse has more than a million records.

Data mining system was created and it was decided to integrate the available data from the e.learning system stored in the common data warehouse (Inmon 2005). To implement this following software was used *IBM* data storage technology *IBM DB2 Universal Database ™ Enterprise Edition, Version 7.2 and IBM's* data mining technology product *Intelligent Miner for Data Version 8.1.*

So transformed data from documental databases enables to perform distributed servers analysis, which consists of the following aspects: user activity assessment, identification of users working with the same databases simultaneously, and perform analysis of e.learning resources.

```
Sub Initialize
                    On Error Goto errorhandler
        // Variables declaration
                    Dim session1 As New NotesSession
                    Dim db As NotesDatabase
                    Dim collection As NotesDocumentCollection
                    Dim doc As NotesDocument
                    Dim st1,st2 As String
                    Dim var As String
                    Dim session As New LCSession
                    Dim conn As New LCConnection("db2")
                    Dim fldList As New LCFieldList
                    Dim fld As LCField
                    Dim i As Long
        ...
                    Set fld7 = fldList.append("CLENGTH",LCTYPE_TEXT)
                    Set fld8 = fldList.append("CTTYPE",LCTYPE_TEXT)
                    Set fld9 = fldList.append("METHOD",LCTYPE_TEXT)
                    Set fld10 = fldList.append("BROWSER",LCTYPE_TEXT)
                    Set fld14 = fldList.append("ETIME",LCTYPE_NUMERIC)
                    Set fld15 = fldList.append("UTRANS",LCTYPE_TEXT)
        ...
                    fld7.Value = doc.contentlength
                    fld8.Value = Strleft(Cstr(doc.contenttype(0)), "/")
                    fld9.Value = Strleft(Cstr(doc.request(0)), "/")
                    fld10.Value = Strleft(Cstr(doc.useragent(0)),"/")
                    fld14.Value = doc.processtimems
                    fld15.Value = Strright(Cstr(doc.uritranslated(0)),":/lotus/domino/data/")
                    Call conn.Insert(fldList)
                                Print  Str$(i)
                                i = i +1
                    Set doc = collection.GetNextDocument(doc)
                    Wend
                    Exit Sub
errorhandler:
                    Dim Msg As String
                    Dim Msgcode As Long
                    Dim status As Integer
                    Dim result As String
                    If session.status <> LCSUCCESS Then
                                status = session.GetStatus(result, Msgcode, Msg)
                    End If
                    Msgbox result
        End Sub
```

Figure 4. The script fragment f proposed agent

Data mining in e.learning system

There are many definitions of data mining, but they are all interrelated and similar. Thus, data mining is the extraction of the un-known or partially known information from the data sets (Baragoin *et al*. 2001; Dunham 2004).

E.learning is relatively new discipline in Lithuania and data mining could be a useful exploration tool in this field. Worldwide e.learning is well known and analyzed. Nevertheless, it remains one of the most interesting fields for the data mining application (Hacid 2004).

Demographic clustering method was used in this analysis. As a rule, clustering is a first step in data mining (Han *et al.* 2001). Once clusters (groups of clusters) have been identified, specific model is created for each of them (Palmer 2000). During result generation, it was decided to create 9 clusters. This method has helped to realize that used activity increases in working days after 4 pm, when users access the public databases. Users frequently access without their identification, i.e. the connection is not to the e-mail or course material, but to the freely accessible information. These can be lecture transcripts, databases of the theses or other web sites of informational nature. Students are mainly from Lithuania. By determining

users working simultaneously with the same databases, the association method was applied. It has helped to identify the probability (in this case even 60%!) for one user to connect to database, if the other user was connected to it as well. The rule can be represented in the following way: *[user] => [courses]*. At the end of the analysis, three most affected databases were identified. They should be distributed among different servers in order to achieve optimal group work. organization and avoid technical problems due to the high number of users.

Using classification method it was revealed that 30% of quereid documents were not returned to the user, and almost always of the same size. With these results it can be concluded that one of the unauthorized users was on-line looking for information, and thus disrupting the servers work. However, it was discovered that such interferences might have been caused by software operated by an individual.

Application of the received results

Recent increase in the web site activity has caused a significant rise in the server load. As a result, queries are often carried out slowly, at peak hours due to the scarcity of resources of the system, server cannot sustain the charge.

Visitor log-in data clearly represents the consumer activity (*X*-axis shows the day of the week, *Y*-axis - the number of log-ins), during almost all the week from Monday to Friday (Fig. 5a). Server load reaches peak between 9am and 10pm. This means that it is the time, when user activity is particularly high. Such user behaviour is a long term trend.

At 9 am the number of users is equal to 45 000, at 10 am - 55 500, at 5 pm - 55 000, at 6 pm - 42 500. Due to the application of data mining technology, the response mode of *gama* server has been modified and the load of the server was lower in the following time slots 10am and 11am, 12am and 1pm and between 2 and 3pm over the week (Fig. 5b). The load of the *gama* server has decreased by 15.4%. Since less user connect to the *irma* server (login is still going through the *gama* server), users are active almost all the week. Peak is reached on Wednesday (13 500 hits), and during the day user activity reaches 5 420 hits at 12 am. *Irma* server user activity is similar to the user activity of *gama* server, and they remain active throughout the day from 9 am till 11 pm. Just like in the *gama* server case, the change in response mode of *irma* server has resulted decrease of load by 5.2% (Fig. 5a., Fig. 5b). Indeed, as shown, user navigation network activity due to traffic and, as a result, it affects Web server performance. To ensure a fast and reliable connection to the server, it is necessary to combine the resources of servers, taking into account the results obtained. Following the interpretation of the research it became possible to reduce server load and manage user queries and hits to the database.

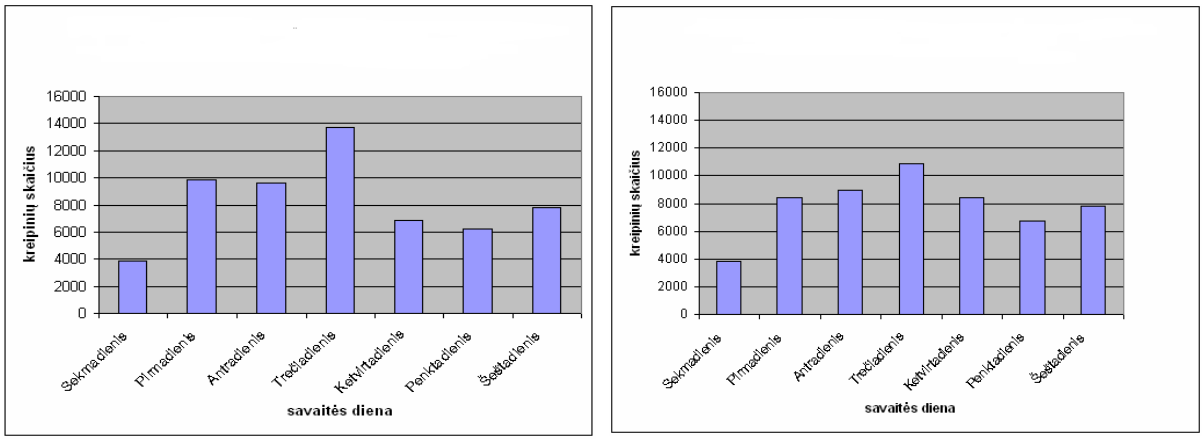

Figure 5. *Gama* server load (*a* – before application of data mining technologies, *b* –after application of data mining technologies)

Summary

Data mining technology allows to identify trends and discover knowledge from the collection of the data. Data mining functions are independent of the choice of the data type: whether it's a huge amount of detailed data on the Internet commerce transactions, or is simply a Web server log-ins. Which data mining function we choose – depends only on solving problem domain.

Data mining technologies applied to the organisation of distributed servers' group work, have allowed to change the servers' response mode. The analysis performed has revealed that server efficiency has substantially increased, especially in case of synchronized work.

Positive results that have been achieved show that models can be used. However, the use of them does not symbolize the completion of data mining process, as these models must be continuously improved. In addition, after the first successful results were obtained, it is important to address the accuracy improvement issues. The methods of the described technology allow to address the real life challenges in the efficient way. At the moment alternative technologies are available not available yet.

It shall be noted that data mining technology is only a decision support systems, and final decisions must be taken by specialist of a field.

The performed research allows the further analysis of e-learning system efficiency and course quality

assessment. In the future it is planned to disclose unauthorized intrusion, to assess e.learning students progress, and start using the *Text Mining* technology to identify student plagiarism, etc.

References

Anjewierden, A., Kolloffel, B., Hulshof, C. (2007). Towards educational data mining: Using data mining methods for automated chat analysis to understand and support inquiry learning processes. *Proceedings of International Workshop on Applying Data Mining in e-Learning (ADML).*

Baragoin, C., Andersen, C. M., Bayerl, S. *et al.* (2001). *Mining Your Own Business in Retail Using DB2 Intelligent Miner for Data*. IBM Corp.

Dunham, M. H. (2003). *Data Mining: Introductory and Advanced Topic.* Prentice-Hall, Upper Saddle River, New York.

Hacid, M. S. (2004). Guest editor's introduction special issue on data mining, *Journal of Intelligent Information Systems (JIIS)*, 22(1), 5–6.

Han, J., Kamber, M., 2001. Data Mining: Concepts and Technigue. Morgan Kaufmann Publisher.

Hanna, M. (2004). Data Mining in the e-learning domain, *Campus-Wide Information Systems*, 21(1), 29–34.

Inmon, W.H. (2005). *Building the Data warehouse.* Fourth Edition. Wiley Publishing, Inc.

Mamčenko, J. (2008). *Duomenų gavybos technolologijų taikymas išskirstytų serverių darbui gerinti.* Daktaro disertacija, Vilnius "Technika".

Palmer, R. C., Faloutsos, C. (2000). Density based sampling: an improved method for data mining and clustering, in *Proc. of the 2000 ACM SIGMOD international conference on Management of data (SIGMOD 2000)*, 82–92.

Srikant, R., Yang, Y. (2001). Mining web logs to improve website organization, in Proc. of the 10th International World Wide Web Conference, Hong Kong.

DUOMENŲ SAUGYKLŲ IR AGENTINIŲ TECHNOLOGIJŲ PANAUDOJIMAS NUOTOLINIO MOKYMO SISTEMOS DUOMENIMS PARUOŠTI DUOMENŲ GAVYBAI

Jelena Mamčenko, Inga Tumasonienė

S a n t r a u k a

Straipsnyje nagrinėjamos netradicinių dokumentinių bazių duomenų paruošimas šiuolaikinėms duomenų gavybos technologijoms. Norint pritaikyti duomenų gavyba nuotolinio mokymosi sistemos duomenis būtina surinkti iš heterogeninių duomenų šaltinių, transformuoti, išfiltruoti ir pritaikyti duomenų gavybos programinei įrangai. Duomenų surinkimas, transformaciją bei integraciją yra atliekami sukurto agento pagalba, kuris gali veikti tiek automatiniame režime tiek gali būti paleidžiamas rankiniu būdu. Pasiūlytas dokumentinių bazių duomenų surinkimo, transformavimo bei filtravimo metodas, buvo realizuotas agentinių technologijų pagalba. Šiame darbe pavyko ne tik pritaikyti duomenis duomenų gavybos programinei įrangai, bet tuo pačiu ir išskirstyti serveriu resursus, perskirstant grupinio darbo organizavimo sistemos duomenų bazes tarp bendradarbiaujančių serverių ir optimizuoti serverių darbą, padidinant jų našumą.

Pirmame įvadiniame skyriuje supažindinama su veikiančia VGTU e.mokymo sistema, išskirstyta tarp bendradarbiaujančių serverių *gama*, *kappa* ir *irma*. Aprašomi serveriai realizuoti *Lotus/Domino* technologijos pagalba, kuri palaiko beveik visus interneto standartus. Aprašoma sistema naudojama jau daugelį metų ir besimokančių studentų kiekis kasmet daugėja, todėl daugėja ne tik duomenų bazių kiekis, bet ir prisijungimų skaičius prie dokumentinių duomenų bazių. Visa tai, savo ruožtų, apkrauna išskirstytų serverių sistemą. Todėl buvo nuspręsta išsaugoti brangius serverių resursus, panaudojant duomenų gavybos technologijas.

Antras skyrius aprašo duomenų gavybos technologijų taikymą dokumentų valdymo sistemoje. Pateiktas grafinis dokumentinių duomenų bazių duomenų integravimas. Aprašoma galimybė pritaikyti ankščiau aprašytas technologijas dokumentinių bazių duomenims. Analizės metu nustatyta, kad beveik visos duomenų gavybos programinės įrangos yra skirtos tradicinėms reliacinėms duomenų bazėms. Siekiant išvengti minėtų problemų ir išsaugoti istorinius duomenis buvo sukurtos atskiros duomenų saugyklos, kuriose buvo patalpinta jau istorine tapusi informacija iš darbinės duomenų bazės. Dėl serverio specifikos vienintelė išeitis buvo sukurti agentą, kuris tiesiai iš dokumentų kolekcijų integruotų ir talpintų duomenis į duomenų saugyklą.

Pirmiausias žingsnis sukurtame agente– parinkti duomenų bazę, iš kurios bus paleistas agentas ir paimti visus dokumentinės bazės dokumentus į dokumentų kolekciją, toliau prisijungiama prie sukurtos *DB2* duomenų saugyklos. Sekančiame žingsnyje kiekvienam laukui priskiriami konkretaus stulpelio parametrai. Priskyrus parametrus vyksta atitinkamos dokumento lauko reikšmės įkrovimas. Šioje vietoje sukurtas ciklas kartojasi, kol pereinami visi dokumentai.
Taip pat atliekamas duomenų valymas.

Naudodami sukurtą agentą, duomenys įdedami į duomenų saugyklą, kurioje yra užtikrintas operacinių duomenų kaupimas, prieš tai atliekant reikalingą duomenų sujungimą, filtravimą ir struktūrizavimą. Informacija apie vartotojus iš *LDAP* serverio, jų aktyvumą, informacija apie kursus ir *log* dokumentinės bazės taip pat yra saugomos duomenų saugykloje. Šiuo metų duomenų saugykloje pateikti kelių mėnesių duomenys iš įvairių duomenų šaltinių.

Trečiame skyriuje trumpai aprašomi duomenų gavybos metodai, kurie buvo panaudoti esamai problemai spręsti ir pateikiami gauti rezultatai.

Ketvirtame skyriuje dėmesys sukoncentruotas į gautų rezultatų interpretaciją. Pateikti ir palyginti duomenys prieš ir po duomenų gavybos technologijų panaudojimo.

Duomenų saugyklų ir agentinių technologijų panaudojimas nuotolinio mokymo sistemai leido ne tik paruošti duomenis duomenų gavybai bet ir pakeisti serverių replikavimo režimą. Remiantis gautais rezultatais, atlikta išskirstytų serverių perkonfigūravimo ir apkrovos analizė leido padidinti serverių darbo našumą ypač sinchroninio darbo metu.

Tolimesniuose tyrimuose planuojama atlikti nuotolinių kursų kokybės analizę, įvertinti nuotolinių studentų pažangumo bei plagiato pasitaikymo galimybę, nustatyti neautorizuotus įsibrovimus ir t.t. Sukurta transformavimo metodika atvėrė plačias galimybes e.mokymo duomenų gavybai.

Jelena Mamčenko, dr., VGTU, docentė. Mokslinių tyrimų kryptis – Informatikos inžinerija (07T). Paskelbusi virš 10 straipsnių tarptautinėse duomenų bazėse referuojamuose periodiniuose mokslo žurnaluose ir kitose leidiniuose. Saulėtekio al. 11, Vilnius. Tel. 2744831.

Inga Tumasonienė, dr., VGTU, lektorė. Mokslinių tyrimų kryptis – Informatikos inžinerija (07T). Paskelbusi virš 10 straipsnių tarptautinėse duomenų bazėse referuojamuose periodiniuose mokslo žurnaluose ir kitose leidiniuose. Saulėtekio al. 11, Vilnius. Tel. 2744831.