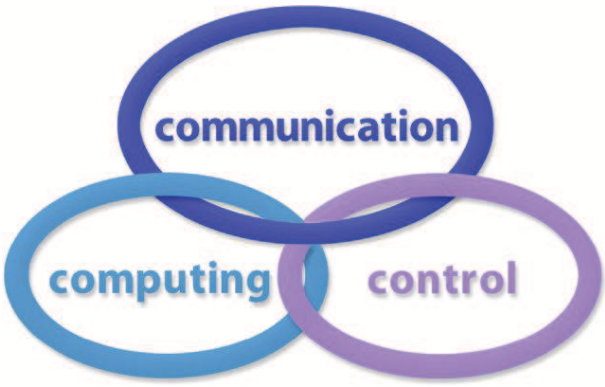
INTERNATIONAL JOURNAL

of

COMPUTERS COMMUNICATIONS & CONTROL

ISSN 1841-9836



A Bimonthly Journal
With Emphasis on the Integration of Three Technologies

Year: 2017 Volume: 12 Issue: 2 Month: April

This journal is a member of, and subscribes to the principles of, the Committee on Publication Ethics (COPE).

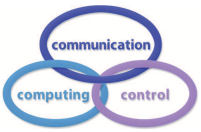
http://univagora.ro/jour/index.php/ijccc/

CCC Publications

Copyright © 2006-2017 by Agora University

BRIEF DESCRIPTION OF JOURNAL

Publication Name: International Journal of Computers Communications & Control.
Acronym: IJCCC; **Starting year of IJCCC:** 2006.
ISO: Int. J. Comput. Commun. Control; **JCR Abbrev:** INT J COMPUT COMMUN.
International Standard Serial Number: ISSN 1841-9836.
Publisher: CCC Publications - Agora University of Oradea.
Publication frequency: Bimonthly: Issue 1 (February); Issue 2 (April); Issue 3 (June); Issue 4 (August); Issue 5 (October); Issue 6 (December).
Founders of IJCCC: Ioan DZITAC, Florin Gheorghe FILIP and Misu-Jan MANOLESCU.
Logo:



Indexing/Coverage:

- Since 2006, Vol. 1 (S), IJCCC is covered by Thomson Reuters and is indexed in ISI Web of Science/Knowledge: Science Citation Index Expanded.

  2016 Journal Citation Reports® Science Edition (Thomson Reuters, 2016):
  Subject Category: (1) Automation & Control Systems: Q4(2009,2011,2012,2013,2014,2015), Q3(2010); (2) Computer Science, Information Systems: Q4(2009,2010,2011,2012,2015), Q3(2013,2014).

  Impact Factor/3 years in JCR: 0.373(2009), 0.650 (2010), 0.438(2011); 0.441(2012), 0.694(2013), 0.746(2014), 0.627(2015).
  Impact Factor/5 years in JCR: 0.436(2012), 0.622(2013), 0.739(2014), 0.635(2015).

- Since 2008 IJCCC is indexed by Scopus (SNIP2014= 1.029):
  Subject Category: (1) Computational Theory and Mathematics: Q4(2009,2010,2012,2015), Q3(2011,2013,2014); (2) Computer Networks and Communications: Q4(2009), Q3(2010, 2012, 2013, 2015), Q2(2011, 2014); (3) Computer Science Applications: Q4(2009), Q3(2010, 2011, 2012, 2013, 2014, 2015).

  SJR: 0.178(2009), 0.339(2010), 0.369(2011), 0.292(2012), 0.378(2013), 0.420(2014), 0.319(2015).

- Since 2007, 2(1), IJCCC is indexed in EBSCO.

Focus & Scope: International Journal of Computers Communications & Control is directed to the international communities of scientific researchers in computer and control from the universities, research units and industry.

To differentiate from other similar journals, the editorial policy of IJCCC encourages the submission of original scientific papers that focus on the integration of the 3 "C" (Computing, Communication, Control).

In particular the following topics are expected to be addressed by authors: (1) Integrated solutions in computer-based control and communications; (2) Computational intelligence methods (with particular emphasis on fuzzy logic-based methods, ANN, evolutionary computing, collective/swarm intelligence); (3) Advanced decision support systems (with particular emphasis on the usage of combined solvers and/or web technologies).

IJCCC EDITORIAL TEAM

Editor-in-Chief: Florin-Gheorghe FILIP
Member of the Romanian Academy
Romanian Academy, 125, Calea Victoriei
010071 Bucharest-1, Romania, ffilip@acad.ro

Associate Editor-in-Chief: Ioan DZITAC
Aurel Vlaicu University of Arad, Romania
St. Elena Dragoi, 2, 310330 Arad, Romania
ioan.dzitac@uav.ro
&
Agora University of Oradea, Romania
Piata Tineretului, 8, 410526 Oradea, Romania
rector@univagora.ro

Managing Editor: Mişu-Jan MANOLESCU
Agora University of Oradea, Romania
Piata Tineretului, 8, 410526 Oradea, Romania
mmj@univagora.ro

Executive Editor: Răzvan ANDONIE
Central Washington University, U.S.A.
400 East University Way, Ellensburg, WA 98926, USA
andonie@cwu.edu

Reviewing Editor: Horea OROS
University of Oradea, Romania
St. Universitatii 1, 410087, Oradea, Romania
horos@uoradea.ro

Layout Editor: Dan BENTA
Agora University of Oradea, Romania
Piata Tineretului, 8, 410526 Oradea, Romania
dan.benta@univagora.ro

Technical Secretary

Domnica Ioana DZITAC
R & D Agora, Romania
ioana@dzitac.ro

Simona DZITAC
R & D Agora, Romania
simona@dzitac.ro

Editorial Address:
Agora University/ R&D Agora Ltd. / S.C. Cercetare Dezvoltare Agora S.R.L.
Piata Tineretului 8, Oradea, jud. Bihor, Romania, Zip Code 410526
Tel./ Fax: +40 359101032
E-mail: ijccc@univagora.ro, rd.agora@univagora.ro, ccc.journal@gmail.com
Journal website: http://univagora.ro/jour/index.php/ijccc/

IJCCC EDITORIAL BOARD MEMBERS

Luiz F. Autran M. Gomes
Ibmec, Rio de Janeiro, Brasil
Av. Presidente Wilson, 118
autran@ibmecrj.br

Boldur E. Bărbat
Sibiu, Romania
bbarbat@gmail.com

Pierre Borne
Ecole Centrale de Lille, France
Villeneuve d'Ascq Cedex, F 59651
p.borne@ec-lille.fr

Ioan Buciu
University of Oradea
Universitatii, 1, Oradea, Romania
ibuciu@uoradea.ro

Hariton-Nicolae Costin
Faculty of Medical Bioengineering
Univ. of Medicine and Pharmacy, Iaşi
St. Universitatii No.16, 6600 Iaşi, Romania
hcostin@iit.tuiasi.ro

Petre Dini
Concordia University
Montreal, Canada
pdini@cisco.com

Antonio Di Nola
Dept. of Math. and Information Sci.
Università degli Studi di Salerno
Via Ponte Don Melillo, 84084 Fisciano, Italy
dinola@cds.unina.it

Yezid Donoso
Universidad de los Andes
Cra. 1 Este No. 19A-40
Bogota, Colombia, South America
ydonoso@uniandes.edu.co

Ömer Egecioglu
Department of Computer Science
University of California
Santa Barbara, CA 93106-5110, U.S.A.
omer@cs.ucsb.edu

Constantin Gaindric
Institute of Mathematics of
Moldavian Academy of Sciences
Kishinev, 277028, Academiei 5
Moldova, Republic of
gaindric@math.md

Xiao-Shan Gao
Acad. of Math. and System Sciences
Academia Sinica
Beijing 100080, China
xgao@mmrc.iss.ac.cn

Enrique Herrera-Viedma
University of Granada
Granada, Spain
viedma@decsai.ugr.es

Kaoru Hirota
Hirota Lab. Dept. C.I. & S.S.
Tokyo Institute of Technology
G3-49,4259 Nagatsuta, Japan
hirota@hrt.dis.titech.ac.jp

Gang Kou
School of Business Administration
SWUFE
Chengdu, 611130, China
kougang@swufe.edu.cn

George Metakides
University of Patras
Patras 26 504, Greece
george@metakides.net

Shimon Y. Nof
School of Industrial Engineering
Purdue University
Grissom Hall, West Lafayette, IN 47907
U.S.A.
nof@purdue.edu

Stephan Olariu
Department of Computer Science
Old Dominion University
Norfolk, VA 23529-0162, U.S.A.
olariu@cs.odu.edu

Gheorghe Păun
Institute of Math. of Romanian Academy
Bucharest, PO Box 1-764, Romania
gpaun@us.es

Mario de J. Pérez Jiménez
Dept. of CS and Artificial Intelligence
University of Seville, Sevilla,
Avda. Reina Mercedes s/n, 41012, Spain
marper@us.es

Dana Petcu
Computer Science Department
Western University of Timisoara
V.Parvan 4, 300223 Timisoara, Romania
petcu@info.uvt.ro

Radu Popescu-Zeletin
Fraunhofer Institute for Open
Communication Systems
Technical University Berlin, Germany
rpz@cs.tu-berlin.de

Imre J. Rudas
Óbuda University
Budapest, Hungary
rudas@bmf.hu

Yong Shi
School of Management
Chinese Academy of Sciences
Beijing 100190, China &
University of Nebraska at Omaha
Omaha, NE 68182, U.S.A.
yshi@gucas.ac.cn, yshi@unomaha.edu

Athanasios D. Styliadis
University of Kavala
Institute of Technology
65404 Kavala, Greece
styliadis@teikav.edu.gr

Gheorghe Tecuci
Learning Agents Center
George Mason University
U.S.A.
University Drive 4440, Fairfax VA
tecuci@gmu.edu

Horia-Nicolai Teodorescu
Faculty of Electronics and
Telecommunications
Technical University "Gh. Asachi" Iasi
Iasi, Bd. Carol I 11, 700506, Romania
hteodor@etc.tuiasi.ro

Dan Tufiş
Research Institute for Artificial Intelligence
of the Romanian Academy
Bucharest, "13 Septembrie" 13, 050711, Romania
tufis@racai.ro

Lotfi A. Zadeh
Director,
Berkeley Initiative in Soft Computing (BISC)
Computer Science Division
University of California Berkeley,
Berkeley, CA 94720-1776
U.S.A.
zadeh@eecs.berkeley.edu

DATA FOR SUBSCRIBERS
Supplier: Cercetare Dezvoltare Agora Srl (Research & Development Agora Ltd.)
Fiscal code: 24747462
Headquarter: Oradea, Piata Tineretului Nr.8, Bihor, Romania, Zip code 410526
Bank: BANCA COMERCIALA FEROVIARA S.A. ORADEA
Bank address: P-ta Unirii Nr. 8, Oradea, Bihor, România
IBAN Account for EURO: RO50BFER248000014038EU01
SWIFT CODE (eq.BIC): BFER

INTERNATIONAL JOURNAL OF COMPUTERS COMMUNICATIONS & CONTROL
ISSN 1841-9836, Volume 12, Issue 2, April, 2017.

Contents

INTERNATIONAL JOURNAL OF COMPUTERS COMMUNICATIONS & CONTROL
ISSN 1841-9836, 12(2):155-165, April 2017.

Workflow Automation in a Risk Management Framework for Pavement Maintenance Projects

D. Benţa, L. Rusu, M.-J. Manolescu

Dan Benţa*
1. IT Center for Science and Technology Bucharest
Romania, 011702 Bucharest, Av. Radu Beller, 25, sector 1
2. Agora University of Oradea
Romania, 410526 Oradea, Piata Tineretului, 8
3. Beck et al. Services Cluj-Napoca
*Corresponding author: dan.benta@bea-services.com

Rusu Lucia
Computer Science for Economics Department
Babes-Bolyai University of Cluj-Napoca
Romania, 400591 Cluj-Napoca, M. Kogălniceanu, 1
lucia.rusu@econ.ubbcluj.ro

Misu-Jan Manolescu
Agora University of Oradea
Romania, 410526 Oradea, Piata Tineretului, 8
mmj@univagora.ro

Abstract: This paper presents a Workflow Management System (WfMS) for procurement process automation in road pavement maintenance and management. It fits information infrastructure for monitoring and maintenance of pavements and roads. Through the two roles of administrator and major users (builder and subcontractors), the solution models the entire process. This way, risks of exceeding allocated budget, time consuming tasks, overcoming deadline, and time consuming quality control, as main issues in risk management, are reduced and controlled.
Keywords: workflow management systems, process automation and control, risk factors, pavement maintenance and control.

1 Introduction

Concerns in terms of pavements design and tests are described in [21] as an action that involves several mandatory steps to assure the best anticipated performance in service. As main steps, authors mention selecting the aggregates, asphalt and additives to be used, testing the asphalt mixtures with varying proportions of the ingredients as close to the field condition as possible, and selecting the optimum mix design. Impacts of geographic location (environment features) and construction type [6] and a flexibility in design, cases when this does not fully respect the standards established [17], has to be considered. Some authors like [16] gave a great importance to socio-economic perspectives and conclude that the direct and indirect environmental, economic, and social impacts, termed as Triple-Bottom-Line (TBL), were not addressed sufficiently. So, one of the main activities in pavement construction projects is the testing phase. The test results influence the material selection for the project. There are several ways to determine the current status of asphalt concretes intended to be used in construction projects, such as core drilling (as destructive method) or Marshall stability test [21] (as non-destructive method). Authors mentioned the Marshall stability test which is related to the tensile strength of the asphalt mixture [21] and [5] and recommended non-destructive testing methods for determining the

Copyright © 2006-2017 by CCC Publications

structural condition of an in-service pavement. Other testing proposed method for determining the material parameters of a fatigue cracking model based on Accelerated Pavement Testing were identified by [5] and [22]. Tests are also essential to determine several features that may influence the time at which the pavement starts to crack, as an essential component of pavement management for planning of maintenance and rehabilitation, where three major features are relevant in terms of the performance of an asphalt pavement [9]: permanent deformation or rutting results, fatigue cracking, and low temperature cracking.

Tests are also relevant for road safety in terms of vehicle handling when tire pavement friction is considered [10]. According to World Health Organization Statistics from May 2016, there are [23]:

- About 1.25 million people die each year as a result of road traffic crashes;
- Road traffic injuries are the leading cause of death among young people, aged 15-29 years;
- 90% of the world's fatalities on the roads occur in low- and middle-income countries, even though these countries have approximately half of the world's vehicles;
- Half of those dying on the world's roads are "vulnerable road users": pedestrians, cyclists and motorcyclists;
- Without action, road traffic crashes are predicted to rise to become the 7th leading cause of death by 2030.

Last stats for Romania according to World Health Organization Statistics are from 2013 and there are 1881 estimated number of road traffic deaths with an estimated road traffic death rate per 100.000 population of 8.7. From 179 analyzed countries, statistics are not encouraging and place Romania on the 5th place, according to the estimated road traffic death rate (per 100 000 population).

Selecting the correct model of pavement is a long term decision as this may affect many adjacent application domains as modeling of the user's path choice behavior, safety, pavement conditions and so on [15].

Many times, materials tracking from request to their use in projects is a difficult task for a real-time overview of supervisors. Besides test results, other factors are essential. Before project start, some parameters like costs, time and quality are defined: a. a budget is agreed and approved, in terms of costs; b. a deadline for completion is set, in terms of time; c. a set of materials and quality (number and special features for each material item) is defined, in terms of quality of work.

The focus of this paper is the development and integration of a WfMS for road pavement maintenance and management (PMMS), according to risk management factors. Section 2 provides several open source solutions for WfMS, part of them can be used as cloud solution and Section 3 presents material resource planning for PMMS, based on JobRouter® WfMS as a cloud computing or on premise offer. Last section formulates conclusions and future work.

2 Related work

The Workflow Reference Model (WfMC) includes five interfaces: Workflow Definition Interchange (Interface 1), Workflow Client Application (Interface 2), Invoked Applications (Interface 3), Other Workflow Enactment Services (Interface 4) and Administration and Monitoring tools (Interface 5) [3].

WfMS provide support in three functional areas [3]:

- build time functions for defining and modeling workflow processes and their activities;
- runtime control functions for managing workflow processes in an operational environment and sequencing activities;
- runtime interactions focused on human users and other application tools for processing the various activity steps.

Other authors joined these three functional areas in two core functional approaches: design time, build time functions and runtime gathering runtime control functions and their interactions. Functional perspective are several parameters: runtime and design time. Runtime parameters are: Research Scope, Installation Time, Documentation, Platform Independent, Easiness of Installation and Utilization, Web Based, Other Software Required, Middleware Platform, DBMS Integration, and Transactions Support. Design time parameters are: Process Definition Time, Documentation, Easiness of the Process Definition Web Based, Organizational Perspective and Workflow Language [19].

We pointed out several open source and Cloud solutions for WfMS, which are most popular and suitable for pavement maintenance.

The YAWL system is an open source workflow solution based on the YAWL (Yet Another Workflow Language) language, and is conformal to the WfMC reference model specifications.

It supports the control-flow perspective, data perspective, and is able to interact with web services. YAWL provides support for workflow patterns, and offers mechanisms that allow persistence, automated form generation and workflow administration. From runtime perspective the installation is simple and it provides a web based application which is user friendly. YAWL requires JRE and Apache Tomcat and is compatible with a middleware platform: SOAP. It allows exceptions treatment during process execution and offers integration with PostrgreSQL as an alternative to Hypersonic, but not with all DBMS compatible. From Design time perspective. Graphical editor provided is not web-based, the definition of own sample process was simple but it does not support the organizational perspective [24].

JOpera offers an administration and monitoring tool, and it implements WfMC reference model. It offers a process definition application, and it is able to interact with other applications. From runtime perspective the system was developed with research purposes, based on Eclipse workbench, and is not a user friendly management environment. JOpera requires Java JDK and Eclipse, provides integration with the most popular DBMS and supports simple exception handling model. From design time perspective JOpera process editor is hardly to use for the definition of process longer, it does not support the organizational perspective. The workflow language is JOpera visual composition language [14].

ProcessMaker is an open source and/or on premise as a cloud solution, which implement WfMC reference model: offers an administration and a client application, offers a process definition application, automate document, approval-based processes across departments and systems. It can be a suitable solution for business users and process experts without programming experience. ProcessMaker contains two main components, a design environment and a run-time engine. The design environment includes tools to map processes, define business rules, create dynamic forms, and manage input and output documents. The run-time engine allows for started and run threads through the process [8, 18].

JBossjBPM is a flexible and extensible WfMS, which implement WfMC reference model: offers an administration/client application, can interact with other applications, and offers a process definition application. From runtime perspective: the installation and usage of this workflow solution quite simple, jBPM requires the installation of JDK and Eclipse with the JBoss IDE plugging, and a middleware platform (Java RMI or CORBA). It offers portability

for most popular databases and supports transactions, allows exceptions treatment and rollback during process execution. From design time perspective JBossjBPM uses BPEL in order to define processes, and the definition of sub processes is not supported. JBossjBPM supports two process definition languages: jPDL and BPEL [11].

Beside those positive features, we also pointed several inconveniences in risk management for pavement maintenance. YAWL and JOpera does not support the organizational perspective, which is the core approach in road maintenance. Due pavement maintenance involve company associations (consortium, collaborative approach) on various sections components, assimilated as sub-processes JBossjBPM is not suitable because the definition of sub processes is not supported.

Another open sources WfMS which have cloud solution are: Bizagi Modeler and a BPM Suite which support the organizational processes and life cycle, Joget Workflow as a web platform for development, deploy and run workflows for organization's business processes, Bonita, Enhydra Shark, JawFlow, JFolder, OpenWFE, RUNA WFE, WfMOpen [8], [19].

3 Workflow automations in PMMS

Risks identified by analyzing the life cycle of a pavement construction project are listed below, ranked according to the conveyed impact, as high, medium or low. High risks factors are: 1. obtaining the necessary funding, 2. exceeding the proposed deadlines, 3. budget over run for various reasons, 4. the deviation from the road axis, 5. instability phenomena, cracks or road crevices appearing in the road embankments, 6. road traffic. Medium risks factors are: 1. lack of personnel, 2. signing a contract with a collaborator, 3. inclement weather (rain, snow), and 4. deviations from the tender book and soils used for embankments. Our workflow intended to decrease high factors (from 1 to 3) and medium factor 2 by implementing an automation workflow for material resource planning, which can be used by companies and collaborators involved in road maintenance projects. Extended risk factors' classification was presented in another paper.

In this case are involved two main actors; user, which can execute own proper tasks according to role and project phase and administrator, which gives users roles and supervises all users and whole workflow of JobRouter®Web frontend offers web-based interface for all user groups (forms for users, reports for process owners or defined users/functions, and workflow designer for process designers). The application allows numerous ways to initiate processes (input as email, files, web service, and scan) and output formats (email, XML/CSV, PDF, logs) activities coordinated by JobServer and JobMail module. The application has a flexible infrastructure and interfaces to any system possible (ERP, file system, CRM, DB, external archive systems, FA) that allows various integration (Figure 1).

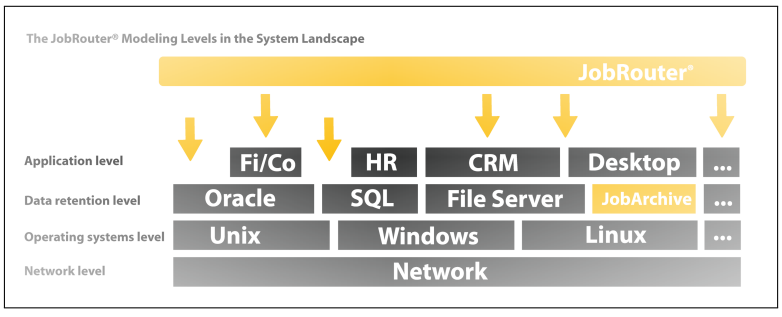


Figure 1: JMultilayer development solution [13]

Multilayer architecture (Figure 2) consists on Workflow Layer, Application Layer (which include interfaces with ERP, CRM, FA), DataStorage for various solution (Oracle, SQL, FS,

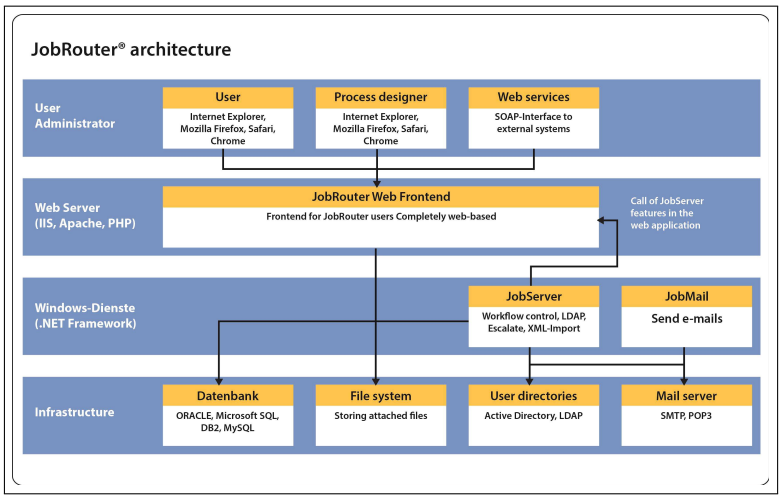


Figure 2: JobRouter® architecture [13]

Archive), and Operating System (Windows or UNIX). It allows users several advantages: each user interaction with the system is stored in the system database, each step can be postponed when this is applicable, each user can request additional information from another user when this is applicable, each generated file is stored and indexed into the application and it can be viewed using the application HTML5 viewer.

With an additional JobTimer module, a deadline can be defined for each process or for each step of the process and with a JobSub module each user can select a substitute for a specific period of time, when not available or on vacation.

Main benefits of the JobRouter® implementation are:

- the application is highly scalable and a stable process engine;
- the application allows multiple system activities for processing and converting data and documents (data maintenance, document conversion and recognition, programs run);
- the application has a fully integrated document archive.

For our main workflow we used: Microsoft Windows Server 2012R2 with Microsoft SQL Server 2014, Internet Information Services (IIS), Visual C++ Redistributable for Visual Studio 2013, Microsoft ODBC Driver 11 for SQL Server, IIS URL Rewrite Module 2.0, and JobRouter® Environment 3.9.5 (then updated to 4.0.0) ioncube_56 install package with JobTable, JobSub, JobPDF, JobSelect and JobArchive modules. The application was tested in all often used browsers and passed all tests for Google Chrome Version 51.0.2704.106 m (64-bit), Firefox 47.0.1, Internet Explorer 11 Version 11.0.9600.17416, Opera 38.0 Version 38.0.2220.41 and Safari 5.1.7 (7534.57.2).

Items appearing once on a material request, as the name of requestor, job function, site, contact details, direct supervisor or date, are stored in a process table with process identification fields as processID, step, status and stepID. Items appearing on several positions on a material request, as the name and characteristic of material, and requested quantity, are stored in a subtable with a subtable view. Several features, as contractors, subcontractors, environment and specific features in PMMS, workers and resources involved, and others, are also defined. This set of parameters materializes in a central database with project details. A centralized record database helps tracking project status and facilitates comparison of what was planned, what was performed and what remains to be improved. All information generated by any user of

the system during the workflow are stored in the database so queries to provide details about the work in a real-time status. Some prerequisites of the system configuration are applicable: when a new user is created in the system, all details as name, function, direct supervisor, site/sites, and contact details are defined; each user is assigned to a function and each function is assigned to a specific step in the workflow; each user can define a substitute for his own user or for his function, when it is not available or on holiday; for each site, relevant details are described (details in terms of costs, completion time and quantity/quality of materials/work); some templates for forms, emails, reminders and generated documents are defined. A CSV procedure is supported by the application, Active Directory Synchronization for application users or single sign on settings to ensure a minimum effort in use of the application.

Designed workflow involves the following steps (Figure 3):

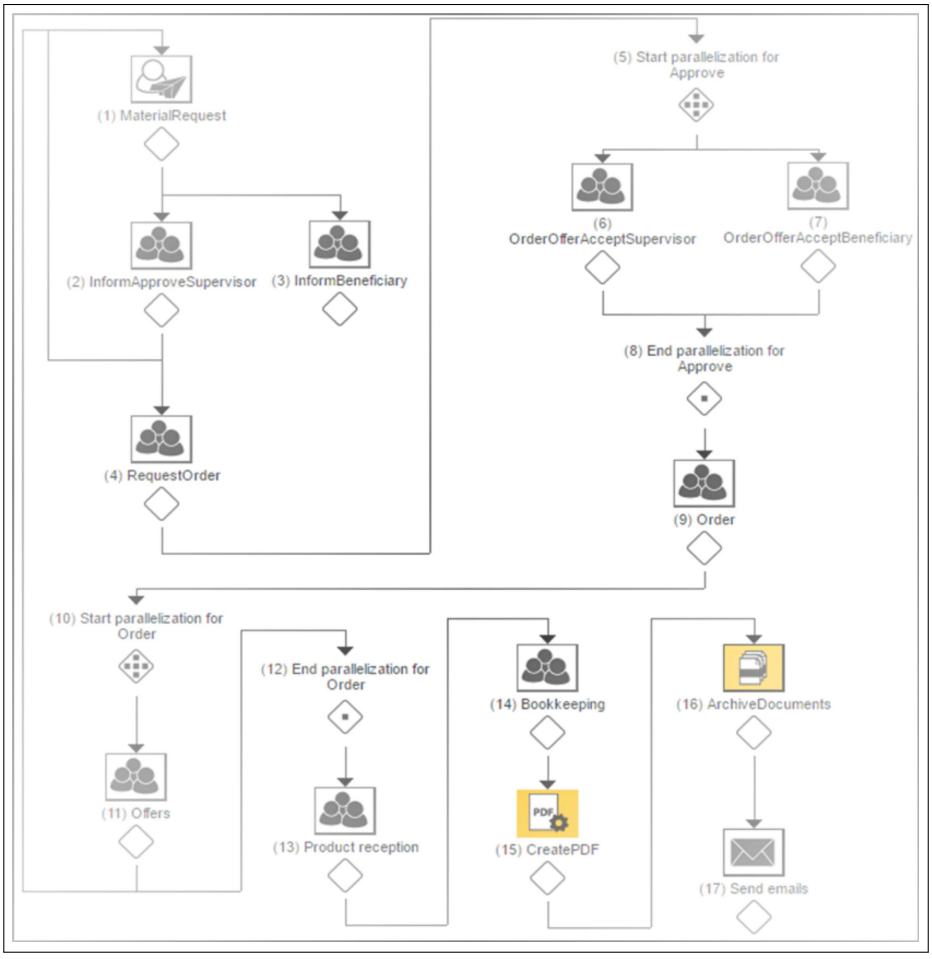


Figure 3: Material Resource Planning Workflow in PMMS

1. MaterialRequest - First step, a worker from a specific site, can start a material request using a standardized form interface. This form contains relevant details about the need and the list of materials; it also automatically retrieves information from the system about the requestor, construction site and other relevant details. The workflow follows then a sequential processing (second step is defined as subsequent step);

2. InformApproveSupervisor - Second step is activated after the first step is processed (the leading step); the supervisor receives the request with all details and he can decline or

accept the request;

3. InformBeneficiary - The beneficiary is informed about the request, too; unlike the supervisor, the beneficiary does not influence the workflow, he is only informed without taking a decision of approving or not;

4. RequestOrder - The request is submitted and send to suppliers; A step can be processed by a user or by a group of users; A processing time is defined (1-3 days in this case to make an offer, depending of each request amount);

5. Start parallelization for Approve - In this parallelization step, the request is sent to the supervisor (site master) and to the beneficiary representative. Each one can approve, reject or request changes for each element of the request. In this parallelization, process is no longer sequential, steps 6 and 7 (starting this point) are simultaneously activated;

6. OrderOfferAcceptSupervisor - The supervisor can approve, reject or request changes for each element of the request;

7. OrderOfferAcceptBeneficiary - The beneficiary representative can approve, reject or request changes for each element of the request. Step (6) and Step (7) can be allocated to different users of functions and can be edited independently. Next step - Step(9) in this case - is activated only after Step(6) and Step(7) are completed;

8. End parallelization for Offer - This is a join step where the status for each element is gathered and approved elements are forwarded in an Order to Step (9);

9. Order - is activated only if both, site master and beneficiary representative, approved the material request and a new order is posted to the selected supplier;

10. Start parallelization for Order - The order is received by selected supplier and requested elements are shipped;

11. Offers - The supplier generates the invoice and each element of the request receives a new status. Some elements are billed and delivered, some of them are not in stock and for some of them similar products are proposed;

12. End parallelization for Order - Each element status is gathered and billed elements are shipped;

13. Product reception - Shipped elements are received by applicant specialized department;

14. Bookkeeping - Details about the order are send to the bookkeeping department for payment and used in future specific accounting procedures;

15. CreatePDF - Request order and clipped documents are stored in PDF files;

16. ArchiveDocuments - Generated documents are automatically indexed and archived in the system for a quick later access. Centralized archive can be web-based accessible and emerging documents are clipped to the initial request order. A detailed overview for each order is accessible;

17. Send mail - When this instance of the process ends, the initiator, the site master and the beneficiary representative are informed that requested materials were received and the bookkeeping department received all necessary documents for payment.

These steps are qualified in: User steps (1 to 10, 9-Order, 13 and 14), Start parallelization step (5, 7) and End parallelization step (8,12) as a beginning and ending sub-process and System Step (15-17). Each involved user can access on-time reports with relevant details about the order and status of order: accepted, declined, changed, received etc. (Figure 4).

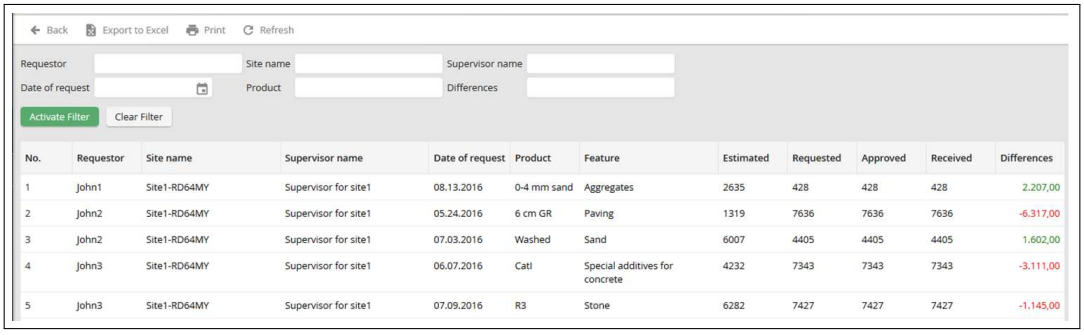
4 Experimental results

Generally, in the maintenance activity is working in teams, each team is coordinated by a manager of the process, called supervisor. There are cases where subcontracting situation often arises; in this case, a number of n companies are coordinated by a manager. The supervisor collaborates with the beneficiary and with the consultant that advises the quality of work. In Figure 3 is represented only the beneficiary because the consultant has no right to decide on material costs.

We selected the most enlightening case study in the process of asphalt (or bitumen) acquisition, used to repair malfunctions. Steps 2 and 3, respectively 6 and 7, are defined as parallel steps because the acquisition must be approved by both representative users, the customer and the process manager.

If case of curbs, fences, traffic signs, marking materials, etc. acquisitions, product acceptance is performed at a specific time t and their use in the maintenance process can be made at time t + n, where n is the number of days elapsed until the real use. For asphalt, if the reception was made, it must be used immediately to repair defective portions because it loses its qualities over time. This situation can lead to a quality risk of the work and time (delays in completion). In this case, step 13 occurs after the real use on the road or sidewalk, without delay. Delivery of the asphalt is made according to the specific needs and risk events occurred. Therefore, steps 9-12 are defined as parallel steps.

Table from Figure 4 shows process status for each material request type. We have 5 columns according to order status and product reception: Estimated, Requested, Approved, Received and Differences. Every step 9 automatically increases Approved values and every step 13 automatically increases Received value.



No.	Requestor	Site name	Supervisor name	Date of request	Product	Feature	Estimated	Requested	Approved	Received	Differences
1	John1	Site1-RD64MY	Supervisor for site1	08.13.2016	0-4 mm sand	Aggregates	2635	428	428	428	2.207,00
2	John2	Site1-RD64MY	Supervisor for site1	05.24.2016	6 cm GR	Paving	1319	7636	7636	7636	-6.317,00
3	John2	Site1-RD64MY	Supervisor for site1	07.03.2016	Washed	Sand	6007	4405	4405	4405	1.602,00
4	John3	Site1-RD64MY	Supervisor for site1	06.07.2016	Catl	Special additives for concrete	4232	7343	7343	7343	-3.111,00
5	John3	Site1-RD64MY	Supervisor for site1	07.09.2016	R3	Stone	6282	7427	7427	7427	-1.145,00

Figure 4: Supervisor report

Differences are presented in suggestive color: red for material and green for suitable differences for materials and are used in risk management analysis as RG=3 for every red values (Figure 4).

Because all messages are sending immediately all steps from 1 to 15, we can have a duration from 5 days to 20 days maximum, depending on suppliers' response. A medium duration is 7 day taking into account 2 weekend days. Moreover, archived documents (step 14) offers possibility to make a complete risk analysis, linked with others risk factors (distress, product events and so on).

Conclusions

Developed application controls project environment risks and, most often, prevent exceeding allocated budget, reduce time consuming tasks and especially improves quality control. Comparing to previous experiences where no IT platforms were used in pavement construction projects, this solution improves all three mentioned parameters: time, funds and quality.

In terms of funds, it creates a centralized database where all material request is stored with related prices. In this way, a real-time tracking for material costs budget lines is performed. In previous cases without IT platform support, there were many cases when budget was exceeded because materials evidence was not available to all sites and reallocation of materials was cumbersome because it was not known if the extra materials on a site could cover the requirements from another site and new orders were performed. Also, there was no high control on costs and on approval chain. In terms of time, it reduces time for orders and approvals, so involved personnel can be relocated to other specific activities of the project. In previous cases without IT platform support, there were many cases when teams were not working because of the materials lack and they were expected materials not yet approved, due to a cumbersome approval and acquisition paper-based procedure. In terms of quality, it improved material quality check and provides support for on-site inspections. This type of application was a time-saving tool for on-site inspectors; it saved time and checked specification criteria when a new on-site inspection was conducted. In previous cases without IT platform support, there were many cases when same verifications were performed by same person because there were no evidence that planned set of materials were ordered and used in the project.

As future work, new workflows for travel in the country or abroad and for life management can be developed. In this way, if a user is away or on vacation, a substitute is automatically activated and respective user will not be included in the construction workflow. Sensors can be integrated to capture relevant data; Collected data can be integrated in several intelligent vehicle systems for user assistance and safety [20] or integrated in web-based systems for routing of pedestrians with different physical abilities within built environments [1] technologies that are also implemented in less developed countries. Concerns are also in terms of safety and security, to embed wireless communications technology on the pavement or surveillance technologies [4], [12] and to collect and transmit different data types, even trigger medical assistance support in case of emergencies [7].

Acknowledgment

This paper was supported by Grant Project Partnerships PCCA2013 "Intelligent management, monitoring and maintenance of pavements and roads using modern imaging techniques-PAV3M" PN-II-PT-PCCA-2013-4-1762, no. 3/2014. We gratefully acknowledge the contribution of colleagues from Beck et al. Services SRL Cluj-Napoca (RO) and JobRouter® AG Mannheim (DE).

Bibliography

[1] Sobek A., Miller H. (2006); U-Access: a web-based system for routing pedestrians of differing abilities, *J GeographSyst* 8: 269-287, DOI 10.1007/s10109-006-0021-1.

[2] Chen W.F., Liew R. (2003); *The civil engineering handbook. New Directions in Civil Engineering.* CRC Press, Boca Raton.

[3] Hollingsworth D. (1995); Workflow Management Coalition The Workflow Reference Model, http://www.bibsonomy.org/bibtex/236043efffa9939b3b1c45a80859207c2/mstrohm

[4] Wright D., (2011); A framework for the ethical impact assessment of information technology, *Ethics InfTechnol*, 13:199-226, DOI 10.1007/s10676-010-9242-6.

[5] Yoo D.G., Kim J., Geem Z.W. (2014); Overview of Harmony Search algorithm and its applications in Civil Engineering, *Evol.Intel.*, 7:3-16, DOI 10.1007/s12065-013-0100-4.

[6] LeeD, Kim S.-K. (2005); Impacts of Geographical Location and Construction Type on As-Built Roughness in Highway Pavement Construction, *KSCE Journal of Civil Engineering*, 9(6): 447-452.

[7] Mordini E., Wright D., Wadhwa K., De Hert P., Mantovani E., Thestrup J., Van Steendam G., D'Amico A., Vater I. (2009); Senior citizens and the ethics of e-inclusion, *Ethics InfTechnol*, 11:203-220, DOI 10.1007/s10676-009-9189-7.

[8] Abdelgader F.M.Z., SalihDawood O.O., Mustafa M.M.E. (2013); Comparison of The Workflow Management Systems Bizagi, ProcessMaker, and Joget, *The International Arab Conference on Information Technology*, 1-5, http://acit2k.org/ACIT/2013Proceedings/158.pdf

[9] Shin H. (2006); Development of a Semi-Parametric Stochastic Model of Asphalt Pavement Crack Initiation, *KSCE Journal of Civil Engineering*, 10(3):189 194.

[10] Oppenheim I., Shinar D. (2012); A context-sensitive model of driving behaviour and its implications for in-vehicle safety systems, *Cogn Tech Work*, 14:261-281, DOI 10.1007/s10111-011-0178-3.

[11] jBPM (2007); JBossjBPM. http://www.jbpm.org/.

[12] Winter J.S. (2014); Surveillance in ubiquitous network societies: normative conflicts related to the consumer in-store supermarket experience in the context of the Internet of Things, *Ethics and Information Technology*, 16(1):27-41, doi:10.1007/s10676-013-9332-3

[13] JobRouter® Optimized BP (2016); Workflow management http://www.jobrouter-workflow.com accessed August 2016.

[14] JOpera Project: Process Support for more than Web Services (2004); http://www.iks.ethz.ch/jopera.

[15] Dimitriou L., Tsekeris T. (2009); Evolutionary game-theoretic model for dynamic congestion pricing in multi-class traffic networks, *Netnomics*, 10:103-121, DOI 10.1007/s11066-008-9027-9.

[16] Kucukvar M., Noori M., EgilmezG., Tatari O. (2014); Stochastic decision modeling for sustainable pavement designs, *Int J Life Cycle Assess*, 19:1185-1199, DOI 10.1007/s11367-014-0723-4.

[17] Pellegrino O. (2011); Road context evaluated by means of fuzzy interval, *Cogn Tech Work*, 13:67-79, DOI 10.1007/s10111-010-0155-2.

[18] PM (2013); Processmaker http://processmaker.com/index.php/ProcessMaker_Architecture_Diagrams2013.

[19] Garcęs R., de Jesus T., Cardoso J., Val P. (2009); Open Source Workflow Management Systems: A Concise Survey, Chapter in Book *2009 BPM & Workflow Handbook*, Publisher: Future Strategies Inc., 179-190.

[20] Bishop R. (2005); Intelligent vehicle R&D: A review and contrast of programs worldwide and emerging trends, *ANN. TELECOMMUN.*, 60(3-4): 228-263.

[21] Terzi S., Karasahin M., Gokova S., Tahta M., Morova N., Uzun I (2013); Asphalt concrete stability estimation from non-destructive test methods with artificial neural networks, *Neural Comput& Applic*, 23:989-997, DOI 10.1007/s00521-012-1023-1.

[22] Suh Y., Mun S., Yeo I. (2010); Fatigue life prediction of asphalt concrete pavement using a harmony search algorithm, *Ksce J CivEng*, 14(5):725-730.

[23] World Health Organization Statistics; `WorldHealthOrganizationStatistics`, accessed August 2016, `http://www.who.int/gho/publications/world_health_statistics/en/`

[24] YAWL, Yet Another Workflow Language, `http://yawlfoundation.org/product/index.php`

INTERNATIONAL JOURNAL OF COMPUTERS COMMUNICATIONS & CONTROL
ISSN 1841-9836, 12(2):166-182, April 2017.

Domain/Mapping Model: A Novel Data Warehouse Data Mode

I. Bojičić, Z. Marjanović, N. Turajlić, M. Petrović, M. Vučković, V. Jovanović

Ivan Bojičić*, Zoran Marjanović, Nina Turajlić, Marko Petrović, Milica Vučković
Faculty of Organizational Sciences, University of Belgrade
Jove Ilića 154, 11000 Belgrade, Serbia
ivan.bojicic, marjanovic.zoran, turajlic.nina, petrovic.marko, vuckovic.milica@fon.bg.ac.rs
*Corresponding author: ivan.bojicic@fon.bg.ac.rs

Vladan Jovanović
Allen E. Paulson College of Engineering and Information Technology, Georgia Southern University
Statesboro, USA
vladan@georgiasouthern.edu

Abstract: In order for a data warehouse to be able to adequately fulfill its integrative and historical purpose, its data model must enable the appropriate and consistent representation of the different states of a system. In effect, a DW data model, representing the physical structure of the DW, must be general enough, to be able to consume data from heterogeneous data sources and reconcile the semantic differences of the data source models, and, at the same time, be resilient to the constant changes in the structure of the data sources. One of the main problems related to DW development is the absence of a standardized DW data model. In this paper a comparative analysis of the four most prominent DW data models (namely the relational/normalized model, data vault model, anchor model and dimensional model) will be given. On the basis of the results of [1][a], the new DW data model (the Domain/Mapping model-DMM) which would more adequately fulfill the posed requirements is presented.
Keywords: data warehouse, data models, relational/normalized model, data vault model, anchor model, dimensional model, domain/mapping model

[a]Reprinted (partial) and extended, with permission based on License Number 4057540167908 [2016] ©IEEE, from "Computers Communications and Control (ICCCC), 2016 6th International Conference on".

1 Introduction

A data warehouse can be defined as a model of a concrete business system representing a set of all of the states of that system during a given interval of time. The constant changes (organizational, legislative, functional, etc.) that a business system faces also reflect on the supporting data warehouse. Hence, one of the main issues, related to DW development and maintenance, is the inconsistency, between the actual system and its supporting data warehouse, which increases over time. Overcoming this discrepancy requires a flexible DW data model i.e. a data model which could be easily adaptable to the frequent changes in the business system.

An additional issue in the field of DW development is the absence of a standardized model for representing the structure of a data warehouse (i.e. a standardized DW data model). Existing approaches propose that the data should be organized in compliance with the third normal form (3NF) [2] or the multi-dimensional model [3]. Both approaches exhibit some limitations related to the difficulty in maintaining the data warehouse when the structure of the data sources changes. On the other hand both approaches are standardized by means of corresponding metamodels defined in the Common Warehouse Metamodel (CWM) [4]. Two additional approaches, aimed at addressing these limitations, have emerged in recent years, namely the Anchor model [5], based on data that has been normalized into the sixth normal form (6NF), and the Data Vault

Copyright © 2006-2017 by CCC Publications

Table 1: Fundamental concepts of the the data models

	Object	**Relationship**	**Attribute**	**Identifier**
Normalized model	Relation	Foreign Key	Domain	Primary Key
Data Vault model	Hub	Link	Satellite	Business / Primary Key
Anchor model	Anchor / Knot	Tie	Attribute	Primary Key
Dimensional model	Dimension	Fact	Attribute	Business / Primary Key

model [6] which can (but need not) also store data normalized into the 6NF. It should be noted that the Anchor and Data Vault models are not standard extensions of CWM.

By identifying the strengths and weaknesses of each of these models it is possible to establish the foundation for a new DW data model which would more adequately fulfill the posed requirements. The comparative analysis of these models is given in [1] (doi: 10.1109/ICCCC.2016.7496754) and based on those results and as an extension, this paper proposes a novel data warehouse data model: the Domain/Mapping Model (DMM).

The remainder of the paper is organized as follows: first the fundamental concepts of the DW data models will be identified and elaborated. Sections 2 is devoted to the analysis of the four most prominent DW data models. Section 3 details the groundwork for the proposed model, which is introduced in Section 4. Several examples illustrating the usage of the proposed DMM are given in Section 5. The final Section gives a brief summary of the work.

2 Comparative analysis of the Data Models

Data models are intellectual instruments for specifying the static characteristics of systems, i.e. for describing the objects, their attributes and their relationships in a stationary state [7]. As a data warehouse is defined as a model of a concrete business system representing a set of all of the states of that system during a given period of time, it is imperative that the underlying data model, be able to, not only support the specification of the system as it transitions through states, but also withstand changes in the business system or data sources.

The four most prominent models are analyzed in this paper: the Normalized model [2, 8, 9], the Data Vault model [6, 10–12], the Anchor model [5, 13] and the Dimensional model [3, 14, 15].

At the highest level of abstraction, all of the described models are based on several fundamental concepts, as depicted in Table 1.

2.1 Built-in semantics

The main point of difference among the models is the level of built-in semantics they provide.

The *Normalized model* does not presume any semantic constraints and, as such is extremely general, as the development of any given business model is based on mappings between sets. Furthermore, it does not provide any implicit concepts which would enable maintaining the history of changes of an object nor the values of its attributes.

The *Data Vault model* assumes that business objects have a stable identifier and somewhat alters the structure of the source by allowing for an object to be arbitrarily structured (i.e. its structure can be split into several *Satellites*). The *Data Vault model* is suited for tracing changes in the values of attributes, except for the *Hub* identifier (i.e. the *Business Key*).

The *Anchor model* is highly normalized. It provides two concepts, the *Anchor* and the *Knot*, for representing business objects. In addition, it enables the tracing of the history of all concepts, save for *Knots*.

The *Dimensional model* is based on the events that take place within a business system and the *Dimensions* which define them. Furthermore, it is possible to define numerical properties for expressing the quantitative aspects of the events. The tracking of the history of changes is based on the complex rules pertaining to changes in dimensions.

All of the models provide a single representation of an object (or entity) except for the *Anchor model* in which concepts of a concrete system can be represented by *Anchor* or *Knot* concepts. The main difference is that the *Normalized*, *Data Vault* and *Anchor models* are normalized, while the *Dimensional model* is denormalized.

A relationship between objects is represented through the *Foreign Key* concept in the *Normalized model*, which establishes a "tight relationship". The *Data Vault*, *Anchor* and *Dimensional models* define the relationship between objects through the *Link*, *Tie* and *Fact* concepts, respectively, wherein the relationship is realized as a table which stores the primary keys of the objects in the relationship. In addition, in *Dimensional model* it is customary to store additional derived or aggregated attributes in the structure of a *Fact*.

With regard to attributes, it should be noted that the *Data Vault* and *Anchor models* separate the structure of an object from the object itself, by using *Satellites* and *Attributes*, respectively, which reference the object via a foreign key. The difference between these two models is that, in the *Anchor model* a separate table is created for each attribute, while, in the *Data Vault model*, the attributes can be grouped according to various criteria into multiple *Satellites* of one object. The *Normalized* and *Dimensional models* keep the attributes within the structure of the object.

In all of the models each concept, which is used for representing an object type, has an identifier. The *Data Vault model* assumes that the identifier is the actual business key. Somewhat similarly, the *Dimensional model*, when using *Type 2 SCDs*, also expects the existence of a business key on the basis of which the dimension values will be grouped.

2.2 Resilience to change

As previously mentioned, the constant changes that a business system faces also reflect on the supporting data warehouse. Hence, one of the primary requirements related to data warehouse development is to provide the ability of absorbing changes in the structure of the data sources, without changing the structure of the underlying data model, in order to facilitate future maintenance and extensions of the data warehouse.

In this section the data models will be analyzed from the viewpoint of their adaptability and extensibility. More precisely, the evaluation of this aspect will be focused on establishing whether changes in the structure of the data sources can be handled simply through the addition of new concepts, without requiring modifications of the DW physical layer.

The *Normalized model* has proven to be the optimal data model for transactional systems since, on the one hand, it guarantees minimal data redundancy and, on the other, decomposes the structure of the system down to the level of its fundamental objects. However, in the data warehouse domain it demonstrates certain weaknesses, as data warehouse requirements differ considerably from transactional system requirements. The main weakness of the *Normalized model* lies in the fact that the attributes and relationships are built into the structure of the system's objects, which leads to a number of maintenance-related issues. Namely, any change in the structure of the source (attributes or relationships) will require changes in the structure of the Normalized model, as depicted in Fig.1, where the cardinality of the relationship between the *Employee* and *Organizational Unit* concepts is adjusted in order to allow for an *Employee* to be assigned to more than one *Organizational Unit*. In the depicted example it is necessary to create the *Position* table, transfer the existing data (pertaining to the relationship between the *Employee* and *Organizational Unit* entities) into the *Position* table, and delete the foreign key,

referencing the *Organizational Unit*, from the *Employee table*.

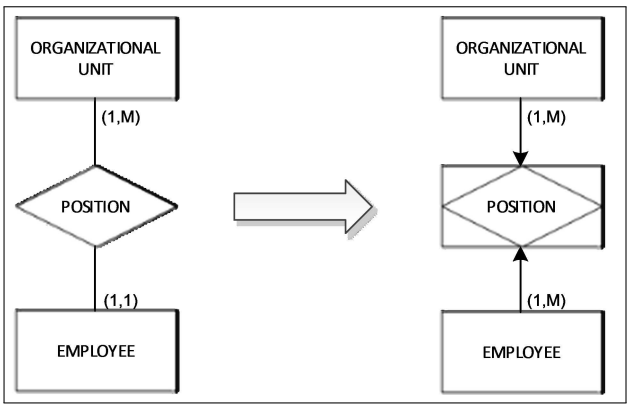


Figure 1: Cardinality change in the *Normalized model*

In addition, the *Normalized model* has some shortcomings when it comes to the reconciliation of source models since, due to the potential semantic heterogeneity of the source models, it may not be possible to design a single reconciled model, as demonstrated by *Golfarelli* in [15].

The *Data Vault model* exhibits much greater flexibility with regard to changes in the structure of the sources. Its flexibility derives from the underlying language, as the structure of an object is decoupled from the object itself, and placed in a physically separate concept: a *Satellite*. Furthermore, every relationship (i.e. *Link*), regardless of its cardinality, is always created as a table representing an aggregation of the related objects.

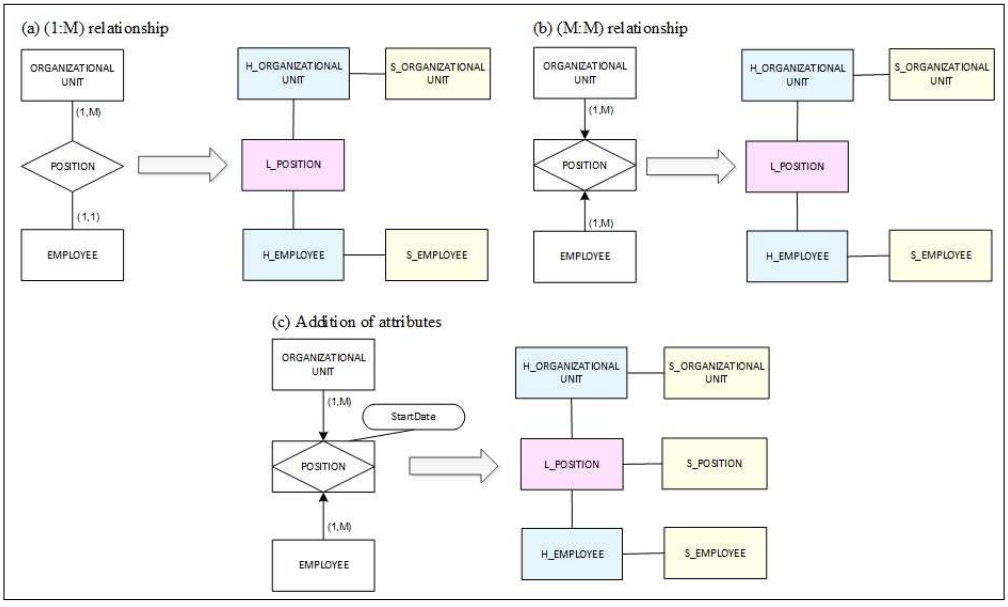


Figure 2: Flexibility of the Data Vault model

As depicted in Fig.2(a), in the initial structure of the source system an *Employee* could be assigned to one and only one *Organizational Unit*. Consequently, the corresponding *DV model* contains the *H_ Employee* and *H_ OrganizationalUnit Hubs* with the associated *Satellites*. The Hubs are related through the *L_ Position Link*. If the structure of the source changes to allow for an *Employee* to be assigned to more than one *Organizational Unit*, the target *DV model* will not undergo any changes, as shown in Fig.2(b).

Furthermore, if the *Position* concept is enhanced with an additional property, e.g. *Start-Date*, the *DV model* once again proves its flexibility, given that it is only necessary to extend (not modify) the existing model by adding an *S_ Position Satellite* (containing the introduced property) to the *L_ Position Link*, as illustrated in Fig.2(c).

Moreover, further modifications, e.g. relating the *Position* concept to other concepts of the source model, will not require changing the existing structure of the *DV model*, given that it allows for two *Links* to be related.

However, it should be noted that, even though in the previous examples the *DV model* did not undergo any changes at the physical level (i.e. it was only extended), the semantics of the source system have implicitly been changed by representing a strong entity (i.e. the *Position* aggregation) as a Link (i.e. event or relationship) in the *DV model*.

An additional drawback of the *DV model* is the fact that, the transformation process is irreversible, i.e. obtaining the structure of the original source model from an existing *DV model* is not possible. Because the *DV model* is semantically much poorer than the source models, certain information contained in the source models will inherently be lost in the transformation. For instance, in the previous examples (Figs.2(b) and 2(c)) it would be impossible to determine into which of the two source models the *DV model* would be reversibly transformed.

In keeping with the same example, the adaptability of the Anchor model, to changes in the structure of the sources, will be examined. All of the models have been created using the original Anchor modeling tool [16].

Initial model (Fig.3(a)) consists of two Anchors: *OU_ OrganizationalUnit* and *EM_ Employee*. The Tie between the two *Anchors* (*OU_ engages_ EM_ assignedTo*), wherein an Employee can be assigned to at most one *Organizational Unit*, is implicitly named on the basis of the roles of the related *Anchors*. Following the transformation the Tie becomes a table whose primary key is the identifier of the concept that participates in the relationship with a maximum cardinality of 1.

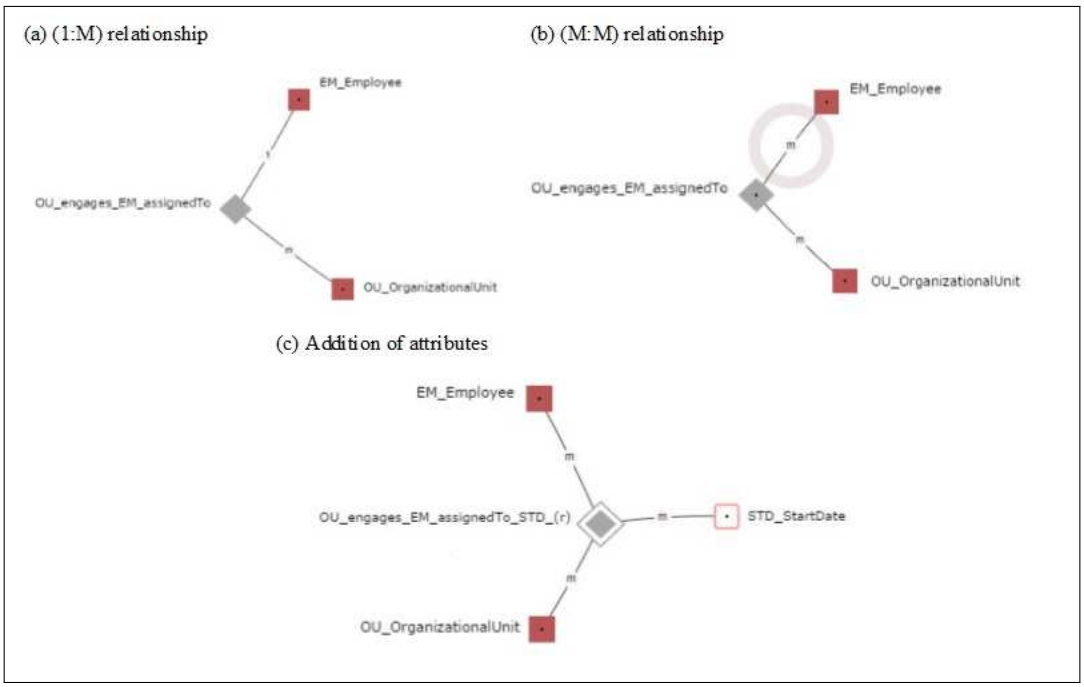


Figure 3: Flexibility of the *Anchor model*

If the cardinality changes, in order to allow for an *Employee* to hold more than one *Position*

(as depicted in Fig.3(b)), the only alteration of the model is extension of the primary key of the *Organizational Unit* with the primary key of the table representing the *Tie*.

However, if the structure of the source system changes so that a *StartDate* is added to the relationship (*OU_ engages_ EM_ assignedTo*), the *Anchor model* does not support this feature. Namely, the transformation of a relationship into an object, or the addition of structural features to a relationship, is not allowed. Instead, the *Tie* can reference a *Knot* representing a set of all of the start dates pertaining to the given *Position*, as depicted in Fig.3(c).

Consequently, at the physical level the table, representing the *Tie*, is modified by adding attribute (the primary key of the *Knot*) which becomes part of the composite key of the *Tie*.

The *Dimensional model* exhibits the same weaknesses as the *Normalized model*, as the addition of new attributes in the source model requires changing the structure of corresponding data warehouse *Dimension*. Yet the relationships in the *Dimensional model* are more stable, given that this model is process-oriented, i.e. it is designed on the basis of the core business processes of an enterprise. Nonetheless, the introduction of additional concepts in the source model (which become *Dimensions* in the data warehouse model) also requires the modification of the corresponding *Fact* in order to include a foreign key referencing the newly formed *Dimension*.

2.3 Temporal aspects

By defining a data warehouse as a collection of Time-Variant data [2], *Inmon* established one of the main requirements for data warehouses, namely that they must enable the tracking of past, current and future states of objects and events, as well as the points in time at which the changes occurred [17]. The notion that a transactional system should support more than one time dimension was first put forward in [18] when *Snodgrass* proposed the *Bitemporal* concept, splitting the time aspect into two dimensions: *Valid time* and *Transaction Time*.

Nowadays, as stated in [19], it is customary for the time aspect to be treated through three dimensions *Valid Time*, *Transaction Time* and *User-Defined Time*. It can be said that a data model supports the modeling of temporal aspects if it provides built-in mechanisms for capturing both the *Valid Time* and the *Transaction Time*.

The aim of this section is to explore whether the data models support the tracking of the entire history of business system objects, and if so, through which mechanisms.

The *Normalized model* does not provide built-in mechanisms for capturing temporal aspects, rather it is left to the data warehouse designer to decide which temporal concepts will be included and how they will be modeled [20].

The *Data Vault model* implicitly incorporates the *Transaction Time* into each concept (*Hub*, *Link* and *Satellite*) via the *Load Date* and *Load End Date* concepts. As this model assumes that the structure of objects will change over time, *Satellites* can be used to capture the changes in values which occur during the life-cycle of an object. Hence the *Valid Time* concept can only be applied to the structure of an object, not the object itself nor its relationships. Consequently, and in light of the underlying premise that business keys are immutable, the *Data Vault model* does not provide a built-in mechanism for tracking the *Transaction Time* for business keys. Nevertheless, such cases can be handled by using the *Same-As Link* whereby two *Hubs*, with different business keys, are asserted to be identical [6].

The *Anchor model* provides the Historized option for representing time-variant *Attributes* and *Ties* which, however, only allows for capturing their *Valid Time*. When this option is used the structure of each of the chosen concepts is supplemented with a *ValidFrom* attribute [21] which captures the beginning of the validity period of the value (e.g. Anchor modeling tool generates a *ChangedAt* attribute for capturing the *Valid Time*). It is implicitly assumed that the end of the validity period of a previous value corresponds to the beginning of the validity period of the

new value. However, such an assumption may not always be valid. In addition, if the Static option is used when creating the model, the possibility of tracking the history of value changes for such *Attributes*, or *Ties*, will be lost. The stated reason is that this option is to be used for values which are considered to be stable, e.g. the *Date of Birth*. Yet, this does not take into account the possibility that the value of an attribute may be incorrectly recorded in the source, or the possibility that several sources, storing different values of the same attribute, may exist. *Anchors* and *Knots* are considered to be immutable, which may not always be the case.

The *Transaction Time* (i.e. *RecordedAt*) is not incorporated into the structure of the concepts representing business system objects, but is stored as metadata in separate tables instead [13].

The Dimensional model enables the tracking of *Valid Time* through the various *SCD types* it offers (excluding *Type 1* and *Type 3 SCDs* which do not allow for tracking the *Valid Time* of an object). However, one of the main issues, related to tracking changes in dimensions, is the choice of the *SCD Type*, as switching to a different *SCD Type* at a later point in time inevitably leads to changes in the structure of the dimension. In addition, if the *Type 2 SCD* is chosen, changes is in the *Business Key* of an object cannot be tracked, as it is the basis for tracking changes in the other attributes.

In the *Dimensional model* the *Transaction Time* is not incorporated into the structure of the concepts representing business objects, instead it is recorded in separate log tables.

2.4 Completeness and traceability of data

In accordance with *Inmon's* definition [2] the second crucial characteristic of data warehouses is the Non-Volatility of data, which assumes that all of the data that has been integrated into a data warehouse must be retained in the data warehouse, unmodified, in order to ensure that a given query, executed at any point in time, will always produce the same, consistent result [22].

The completeness and traceability requirements are in direct collision with the *Single Version of the Truth* concept which assumes that the data, that is to be stored in the data warehouse, is prepared and filtered in advance. The *Single Version of the Truth* is, thus, the result of integrating data from multiple sources with the aim of providing a uniform basis for analysis and avoiding redundancy [23]. However, in order to achieve this aim, it is customary when designing a data warehouse, to choose, out of all of the data relating to a single concept and extracted from various sources, which one will represent the "truth" and, as such, be stored in the data warehouse. All other occurrences (i.e. those which do not conform to the "truth") are discarded and will not be loaded into the data warehouse. As a result, given that the data warehouse stores only part of the source data, it is impossible to conduct analyses on all of the values which exist in the source systems.

Consequently, the *Single Version of the Facts* concept emerged which assumes that the data warehouse stores all of the source data [24], thereby making it possible to provide different views for different users according to their specific needs. In essence, this approach promotes an *ELT* (*Extract-Load-Transform*) process in which the transformation of data takes place after the data warehouse has been loaded, as opposed to the traditional ETL approach in which the data is transformed prior to its being loaded into the data warehouse. As a result, the data warehouse will store all of the available data from all of the sources spanning the entire history of the business system. Therefore, the analytical processing of the data will be performed on-demand depending on the business users needs [24].

The aim of this section is to explore whether the data models provide mechanisms for, on the one hand, ensuring the completeness of the data which is transferred from the sources and, on the other, for tracing the stored data back to the sources from which it originated.

The *Normalized model*, as was the case with the temporal aspect, does not implicitly provide

Table 2: Comparison of DW Data Models

	Normalized	**Data Vault**	**Anchor**	**Dimensional**
Built/in semantics	No	Yes	Yes	Yes
Resilience to change	No	Partially	Partially	No
Temporal aspect	No	Partially	Partially	Yes
Completness	No	Yes	Partially	No
Traceability	No	Yes	Yes	No

concepts to support the traceability of data or the recording of data originating from multiple sources. Traceability can, for example, be accomplished by recording additional metadata for each n-tuple, if the *Single Version of the Truth* approach is adopted. However, if the *Single Version of the Facts* approach is adopted the *Normalized model* demonstrates certain weaknesses. Namely, storing data from several sources in the same model would require storing one n-tuple per source in which the same business concept is present. This leads to the issue of relating these n-tuples, i.e. it requires using additional concepts for relating n-tuples representing the same business object. Furthermore, it introduces redundancy (as the n-tuples representing the same business object contain a number of attributes with the same values) thereby eliminating one of the good traits of the *Normalized model*.

The *Data Vault model* supports the traceability of data by requiring that the source, from which the *Hub* was initially loaded, be recorded. Likewise, the *Satellite* and *Link* concepts include built-in attributes for recording the source from which the loaded values originated (i.e. the *RecordSource* attribute) [11]. However, the main shortcoming of this model lies in the fact that the structure of a Hub contains the business key, the value of which is assumed to be stable or rarely changes. Moreover, the structure of the business key might also change. These situations are resolved by introducing a new *Hub* which will be related to the original Hub via a *"Same-As"Link*. Consequently, more than one *Hub*, with the same attributes and the same relationships with other model elements, will exist. In addition, the *Data Vault model* enables tracing data from multiple sources, by recording the data source in a *Satellite*.

The *Anchor model* supports the traceability of data, via the metadata concept, as all time-variant concepts (thus excluding *Anchors* and *Knots*) can reference the metadata capturing the source of the data. In effect, this also means that it is not possible to record that, for a single object two identical values, originating from two different sources, are both valid at the same point it time. Reason for this is that, at the implementation level, there exists a *Unique Constraint* over a group of attributes: identifier, *Valid Time* and the value of the attribute.

The *Dimensional model* does not provide built-in mechanisms for tracing the stored data back to the sources. Furthermore, it does not allow for recording multiple values (originating from multiple sources) for a single object, during the same period of time.

3 Groundwork

In light of the previous discussion it can be concluded that none of the analyzed models completely fulfill all of the necessary requirements pertaining to DW data models (as summarized in Table 2). By defining and studying the issues, recognized in the previous section, a number of conclusions will be made which will set the grounds for designing a new DW data model.

Extensibility and adaptability of the data model

As previously demonstrated, data models with a higher degree of integration among objects and attributes or objects and relationships (i.e. the *Normalized* and *Dimensional models*) exhibit

less flexibility and adaptability to changes in the structure of the data sources in comparison to models with a lower degree of integration (i.e. *Anchor* and *Data Vault models*).

Example 1: The illustrative examples given in Section 2.2.

Conclusion 1: The data model should reinforce a "loose" coupling between the objects and their attributes and relationships, and support the automated transformation of source model concepts into the 6NF.

Resilience to changes in the values of object identifiers

Although this issue was not explicitly described in the previous analysis of data models (because the identifier is in fact part of the structure of an object), it is relevant because the analyzed models differ in the way that they handle object identifiers. An object identifier is defined as an attribute with a unique value where the inverse *Domain→Object mapping* also has a (0 or 1:1) cardinality. While such a viewpoint is justified in transactional systems, it cannot be applied to the business identifiers of objects when it comes to temporal systems with, potentially, several data sources.

Example 2.1: The Pension and Disability Insurance (PIO) Fund of the Republic of Serbia, established by the Law on Pension and Disability Insurance has two, in the most part, independently maintained information systems in its Belgrade and Novi Sad branches. Consequently, there are a number of cases in which the two branches assigned two different valid *Personal Numbers* (representing business identifiers) to the same person. Given that the integration of these two information systems requires the storing of both identifiers, the same object will have two different, yet simultaneously valid, values for the same identifier. On the other hand, in several cases a single *Personal Number* was assigned to more than one person.

Example 2.2: The Ministry of Interior of the Republic of Serbia is responsible for assigning *Unique Master Citizen Numbers* (in Serbian: *Jedinstveni Matični Broj Gradjana*- JMBG) to all citizens. However, there are some cases in which a citizen was mistakenly assigned two different JMBG numbers which are both valid at the same time. In addition there are a few cases of duplicate JMBG numbers, i.e. the same JMBG was assigned to different citizens.

Conclusion 2: The data model should provide for defining an object identifier, but also allow for a 1:M cardinality not only for the *Object→Domain mapping*, but for the inverse *Domain→Object mapping* as well. In other words, object identifiers should be attributes with multiple, possibly shared, values. The same holds for all other attributes of an object.

Resilience to changes in the structure of object identifiers

The structure of an object can, over time, undergo changes with regard to the attributes (or group of attributes) representing the business identifier. This issue is related to those models which presume the existence of a business identifier (the *Data Vault* and *Dimensional models*).

Example 3.1: Up till 1982, the business identifier of an *Insured Person* object in the PIO Fund of the Republic of Serbia was the *Personal Number*. With changes in legislation, JMBG were introduced as object identifiers.

Example 3.2: Up till 2003, the business identifier of a *Contributor* object in the PIO Fund was the *Registration Code*. With changes in legislation, a group of attributes was introduced as a composite object identifier: the *Tax Identification Number* (in Serbian: *Poreski Identifikacioni Broj* PIB), *Municipality* and *Street*.

Conclusion 3: The data warehouse data model should enable storing semantically different business identifiers for a single object in different periods of time.

Data redundancy

Even though most of the analyzed data models (save for the *Dimensional model*) dedicate special attention to this issue, data redundancy will inevitably occur as new data sources are added to the data warehouse at various points in time. This issues stems from the fact that all of the models structurally relate the attributes to their corresponding objects.

Example 4: The information system of the Ministry of Interior of the Republic of Serbia has, among many others, two subsystems: *Register of Births* (responsible for assigning and processing *Unique Master Citizen Numbers*) and *Human Resources*. When the first subsystem data is loaded, which ever one that may be, the JMBG will, as an attribute, become part of the object it belongs to (e.g. an *Individual*). The loading of the second subsystem data (e.g. *Human Resources* and the *Employee* object which also contains a JMBG attribute) will lead to the storing of the same JMBG values in two different places, independent of one another, in the same DW.

Conclusion 4: The domains, i.e. the sets from which the attributes of an object take their values, should be structurally independent, so as to allow for the same values to be used by the attributes of different objects in order to keep the data redundancy as minimal as possible.

Capturing the temporal aspect

As discussed in Section 2.3, the capturing of temporal aspects is only partly supported in most of the analyzed data models.

Example 5: The information system of the Customs Administration, which is a part of the Ministry of Finance of the Republic of Serbia, has, among others, two subsystems: the *Transit system* (*New Computerized Transit System* - NCTS) and *Reference Data* RD. The NCTS subsystem regularly relies on the RD codelists. The RD codelists have a validity period, and new codelists are introduced into the system at least one month before the beginning of their validity period. It is obvious that all of the system's objects are subject to the temporal aspect.

Conclusion 5: The data model should allow for capturing the temporal dimension, i.e. the valid time and transaction time for every object, attribute and relationship within the system. Furthermore, the capturing of the temporal aspect should be implicit, i.e. it should not depend on the expertise of the designer nor the degree of knowledge about the real system.

Capturing the temporal aspect

This issue was described in Section 2.4. While none of the analyzed data models were, for the most part, initially designed to satisfy this requirement, the *Data Vault* and *Anchor models* do exhibit certain flexibility due to the fact that attributes and relationships are not structurally integrated into the objects. However, the shortcomings of these models are exposed when the temporal dimension is introduced. Namely, when the source model is replaced by another model version the mechanisms provided by the two data models do not allow for tracking the continuity of the two sources, i.e. they cannot be perceived as two versions of the same model.

Example 6: It has been illustrated, through several examples, that it may be expected for the data warehouse to store several different, yet simultaneously valid, values for the same attribute.

Conclusion 6: The data model should implicitly allow for simultaneously storing several different values for a single object characteristic (relationship or attribute) while also maintaining a reference to the version of the source (model) from which the value was retrieved.

All data is equal

Example 7: Two of the analyzed models, namely the *Data Vault* and *Anchor* models, introduce novel concepts which differ semantically from the concepts which are customarily used for describing business objects. The *Data Vault model* provides the *Reference data* concept, while the *Anchor model* offers the *Knot* concept. Both concepts play a role in representing static, immutable objects.

The Ministry of Health of the Republic of Serbia maintains various sets of objects related, for example, to the organizational structure, human resources or medical equipment of the health facilities within its jurisdiction. The processing of such objects relies on the reference data which belongs to a separate *CLASSIFICATIONS* submodel, which represents a unique codelist system accessible by all business objects.

Such codelists would be represented as *Reference data* in *Data Vault* models or *Knots* in

Anchor models. The main limitation of such an approach is that codelists are regarded as immutable, even though they are managed by another business function, i.e. they are managed, like any other object in the system.

Conclusion 7: The data model should give equal importance to each type of data, regardless of whether it represents metadata, whether it originates from within the business system or the frequency with which it changes.

4 The Domain/Mapping model

Taking into account the issues and conclusions described in the previous section, a new model - the **Domain/Mapping model** (DMM) is proposed. The proposed DMM, depicted in Fig.4 is a general model which has been specifically designed to reconcile the semantic differences of existing data warehouse conceptual models, eliminate redundancy, be resilient to changes, allow for the capturing of temporal aspects, enable traceability, extensibility and adaptability, as well as to maintain a single version of the facts.

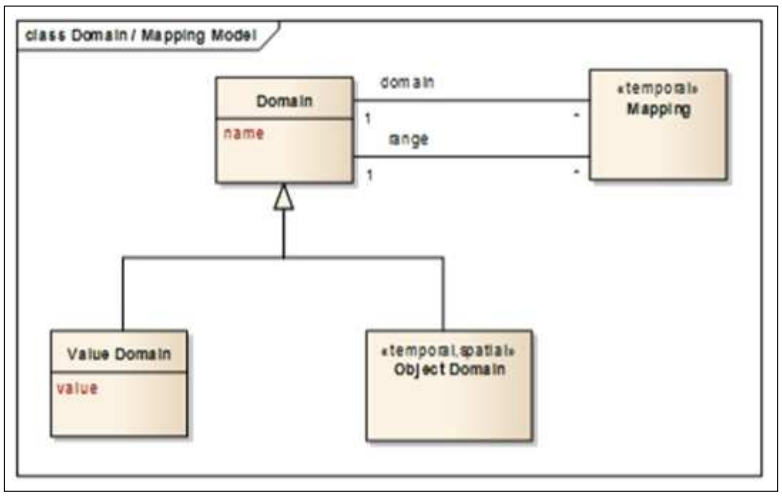


Figure 4: The Domain/Mapping model conceptual model

The core concept of the proposed model is the *Domain* which represents a set of values or objects. A set of values represents a *Value Domain* which is universal and does not depend on temporal and spatial aspects. It is a set of a finite number of atomic elements, from which the attributes of an object take their values. Each such set is predefined to include an additional element: the unknown (i.e. *null*) value.

Definition 1. A Value Domain $D_v = (I, V)$ is an ordered pair, where:

- I is a finite set of identifiers and
- V is a finite set of values.

A value domain is represented by an ellipse symbol with the name of the value domain inside the ellipse.

An *Object Domain* represents a set of real-world objects which are influenced by temporal and spatial aspects, i.e. the objects are mutable in time and space.

Definition 2. A Object Domain $D_o = (I, T, S)$ is an ordered triplet, where:

- I is a finite set of identifiers,

- T is a temporal indicator and
- S is a spatial indicator.

An object domain is represented by a double-lined ellipse symbol with the name of the object domain inside.

A *Mapping* is a concept which allows for forming the structure of the objects and their relationships. When an attribute is defined, then the mapping is between an *Object Domain* and a *Value Domain*. If a relationship between objects is defined, then the mapping is between the *Object Domains*. The forming of the structure of the objects and their relationships is influenced by the temporal aspects. Furthermore, the data warehouse structure is defined in accordance with a particular model version (different sources or versions of the same model).

Definition 3. A Mapping $M = (F, T, V)$ is an ordered triplet, where:

- M is the mapping between two domains,
- T is a temporal indicator and
- V is the version of the model in which the mapping is defined.

A mapping is represented by a solid undirected line with the name of the mapping on it.

Definition 4. F is a pair consisting of a mapping function i.e. a mapping $f(x) = D_d \rightarrow D_r$ and its inverse mapping $f\prime(x) = D_r \rightarrow D_d$, where:

- D_d is the domain of the mapping and,
- D_r is the codomain (i.e. range) of the mapping.

Definition 5. Mappings between two value domains are forbidden.

Definition 6. A Temporal Aspect $T = (T_{tt}, T_{vf}, T_{vt})$ is an ordered triplet of temporal dimensions, where:

- T_{tt} is the transaction time,
- T_{vf} is the beginning of the validity period and
- T_{vt} is the end of the validity period.

Definition 7. A Spatial Aspect $S = (L_t, L_g)$ is an ordered triplet of temporal dimensions, where:

- L_t is the latitude and
- L_g is the longitude.

Definition 8. A Latitude $L_t = (G_d, G_m, G_s, G_{dw})$ is an ordered quadruple, where:

- G_d = the degree ranging from 0-90,
- G_m = the minutes,
- G_s = the seconds and
- G_{dw} = the direction (north/south).

Definition 9. A Longitude $L_g = (G_d, G_m, G_s, G_{dh})$ is an ordered quadruple, where:

- G_d = the degree ranging from 0-180,
- G_m = the minutes,
- G_s = the seconds and
- G_{dh} = the direction (west/east).

5 DMM examples

In this section several examples, demonstrating the utilization of the proposed DMM model, will be given. The depicted examples also illustrate the aptness of the proposed model with regard to the issues and the fulfillment of the requirements postulated in Section 3.

Two simplified submodels of the PIO Fund model, depicting changes in the structure of the *Insured Person* concept during a given period of time, are shown in Fig.5. As depicted, the initial *Insured Person* concept was later modified by introducing the JMBG, which then also became the business identifier of the object due to changes in legislation. Up till then the business identifier was the *Personal Number* (PN). The fulfillment of ***Conclusions 1-3*** will be demonstrated through this example.

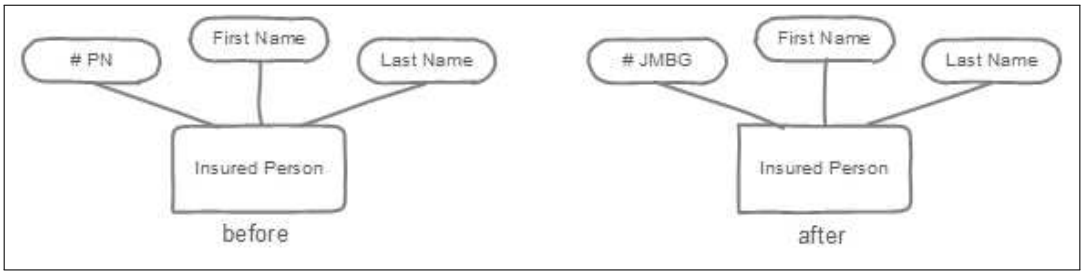


Figure 5: Business identifier replacement

The following Fig.6(a) depicts the DMM model with three domains: *Personal Numbers*, *First Names* and *Surnames*. It also contains the *Insured Person* concept, whose concrete instances are related to the defined domains, via the *Mapping* concept.

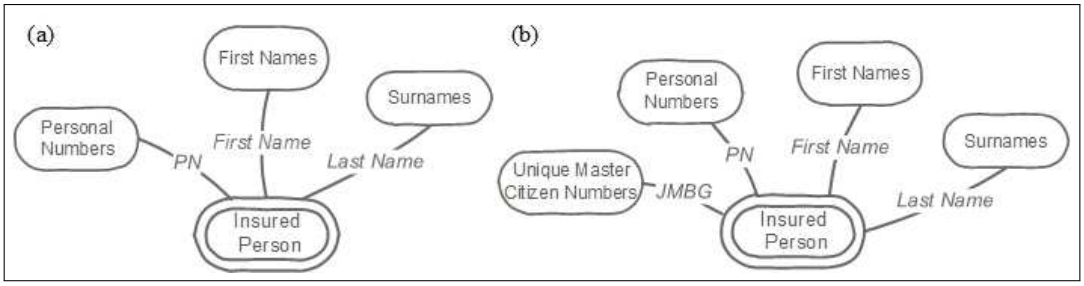


Figure 6: The initial (a) and modified (b) *Insured Person* DMM model

All mappings have an M:M cardinality (i.e. all attributes are multi-valued) which satisfies the requirement that several values of the same attribute can be stored at the same time, across two dimensions: the temporal and source dimensions. Namely, if several different systems store different values of a single attribute, and if the value of the attribute may change over time, the proposed model would be capable of upholding these changes. This is possible due to the

fact that the mapping of domains is accomplished through the *Mapping* concept which implicitly stores both the valid time and the version of the source metamodel. Consequently, the data warehouse stores all values of all data source model versions, without loss of information.

Fig.6(b) illustrates this scenario, when a new attribute, which also becomes a new object identifier, is added. At the physical level, all of the introduced concepts are implemented as tables. As depicted in Fig.6(b), the initial model was modified by simply adding new structures, without altering or deleting existing ones, i.e. by mapping corresponding domains.

The next example demonstrates how the elimination of data redundancy is achieved by keeping the domains (from which the attributes take their values) structurally independent, thereby allowing for several different attributes to use the defined domains at the same time. The depicted example shows that different business functions from different data sources use the same value domains. Fig.7 depicts a simplified *HUMAN RESOURCES* submodel (of the model of the Ministry of Interior of the Republic of Serbia) which contains an *Employee* concept consisting of the *JMBG*, *First Name*, *Last name* and *Birth Date* attributes. The first three attributes are mappings onto value domains, while the *Birth Date* is a mapping onto an object domain *DATE*, which is itself composed of three value domains: *Days*, *Months* and *Years*.

If a new data source is introduced, it is possible to reuse the existing domains, in order to eliminate redundancy, without altering the existing data warehouse structure. The following Fig.7 depicts a *REGISTER OF BIRTHS* submodel of the same business system, which describes the part of the system related to the JMBG of citizens. This submodel maintains the *JMBG* numbers of all citizens and includes all of the JMBG numbers from the human resources department, i.e. those belonging to the employees of the Ministry of Interior. It is customary for those *JMBG* numbers to be stored redundantly, which leads to data anomalies. The redundancy is eliminated by mapping both concepts (*Employee* and *Individual*) to the value domain *Unique Master Citizen Numbers*.

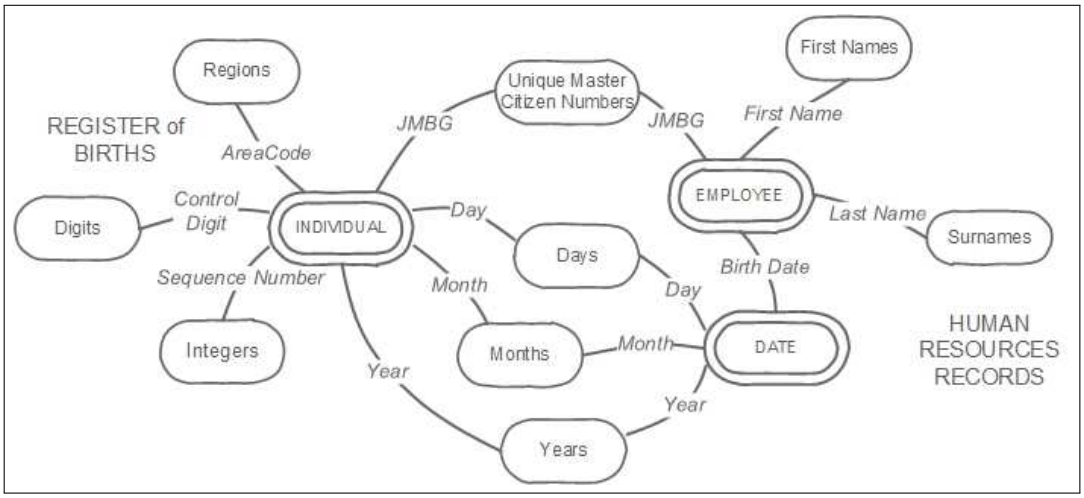


Figure 7: The absence of data redundancy in the DMM model

Furthermore, the Individual concept also reuses the *Days*, *Months* and *Years* value domains, as the concrete values of these domains are utilized for constructing the *Unique Master Citizen Numbers* (the first 7 digits of the JMBG correspond to the date of birth of an individual: two digits for the day, two for the month and the last three digits of the year of birth). By building a net of domains, which can be used for various concepts, data redundancy is kept minimal, which addresses ***Issue 4*** described in Section 3.

In addition, the DMM fulfills the requirement postulated in ***Conclusion 5***, by implicitly

capturing the valid time and transaction time for all objects stored in the data warehouse, as well as for the mappings they participate in.

The fulfillment of the requirement postulated in ***Conclusion 6***, is accomplished by using an implicit M:M cardinality for object-attribute or object-object mappings. Consequently, it is possible for a single object to have several different values for the same attribute or to be simultaneously related to several objects of the same type.

Finally, the fulfillment of the requirement postulated in ***Conclusion 7***, is depicted by the following two submodels (Fig.8). The first shows a simplified *MEDICAL EQUIPMENT* submodel (which is part of the Ministry of Health model).

If the *Equipment Codelist* is modeled without the capability of tracing changes in its values through time, it would not be possible to maintain the classification model without changing its structure. The following submodel *CLASSIFICATIONS*, shows how a new source can be easily introduced into the data warehouse without requiring changes in the existing model, even if it the two models are not at the same level of abstraction.

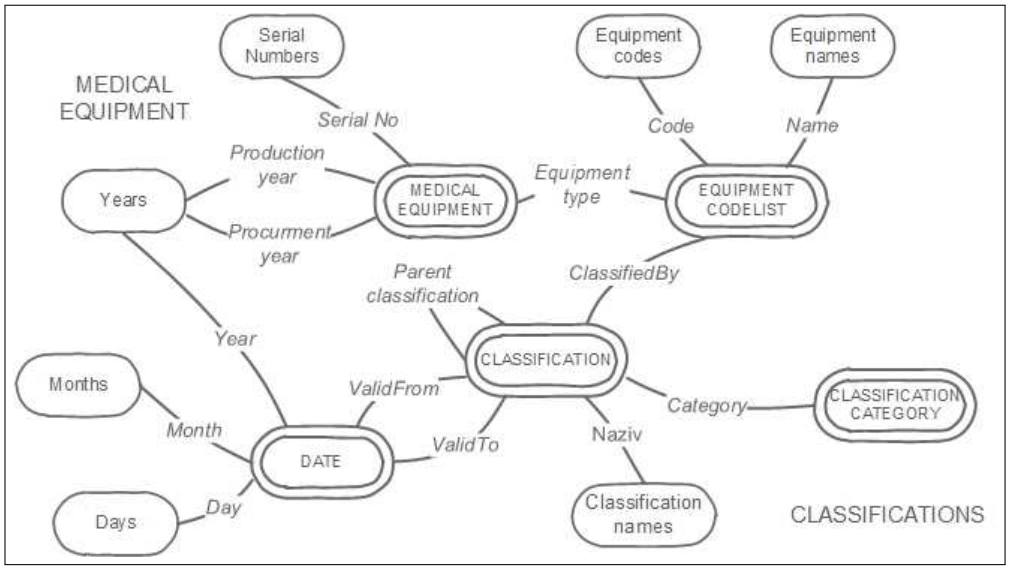


Figure 8: Medical equipment

As demonstrated by the previous examples, a DMM consists of two fundamental concepts, *Domains* and *Mappings*, which provides complete flexibility at the physical level while overcoming the weaknesses of the presented data warehouse data models.

Conclusion

The proposed DMM is capable of absorbing the changes which may occur in the source structures, trace the data back to their sources, and implicitly keep track of both temporal and spatial dimensions, while at the same time providing an integrated view of the data stored in a DW.

In addition, given that the DMM is a general model it can reconcile the semantic differences of the conceptual models used for describing data warehouses, and is, as such, completely independent of the underlying conceptual model.

This opens up the possibility for utilizing the DMM for building the data warehouse modeled by most of the existing conceptual models.

Bibliography

[1] I. Bojičić, Z. Marjanović, N. Turajlić, M. Petrović, M. Vučković and V. Jovanović (2016), A comparative analysis of data warehouse data models, *Computers Communications and Control (ICCCC), 2016 6th International Conference on*, IEEE Xplore, e-ISBN 978-1-5090-1735-5, doi: 10.1109/ICCCC.2016.7496754, 151-159.

[2] W. Inmon, *Building the Data Warehouse*. Wiley, 2002.

[3] R. Kimball, L. Reeves, M. Ross and W. Thornthwaite (1998), *The Data Warehouse Lifecycle Toolkit: Expert Methods for Designing, Developing, and Deploying Data Warehouses*, Wiley, 1998.

[4] Object Management Group (2002); *Common Warehouse Metamodel*, available at: `http://www.omg.org/cgi-bin/doc?formal/03-03-02.pdf`

[5] O. Regardt, L. Ronnback, M. Bergholtz, P. Johannesson and P. Wohed (2009); Anchor Modeling: An Agile Modeling Technique Using the Sixth Normal Form for Structurally and Temporally Evolving Data, *in Proc. of ER09 (Brazil)* , LNCS, 5829(1): 234-250.

[6] D. Linstedt (2010); *Data Vault Modeling Specification v1.0.9.*, available at: `http://danlinstedt.com/datavaultcat/standards/dv-modeling-specification-v1-0-8/`

[7] B. Lazarević, Z. Marjanović, N. Aničić and S. Babarogić (2010). *Baze podataka*. Fakultet organizacionih nauka, 2010 (Textbook in Serbian).

[8] E. F. Codd (1969); Derivability, Redundancy and Consistency of Relations Stored in Large Data Banks, *IBM Research Report*, San Jose, California, 1969.

[9] E. F. Codd (1970); A Relational Model of Data for Large Shared Data Banks, *in Communications of the ACM*, 13(6):377-387.

[10] D. Linstedt (2002); *Data Vault Series 1 - Data Vault Overview*, available at: `http://www.tdan.com/view-articles/5054/`

[11] D. Linstedt (2011); *Super Charge Your Data Warehouse: Invaluable Data Modeling Rules to Implement Your Data Vault*, CreateSpace Independent Publishing Platform, 2011.

[12] D. Linstedt (2003); *Data Vault Series 2 - Data Vault Components*, available at: `http://www.tdan.com/view-articles/5155/`

[13] L. Ronnback, O. Regardt, M. Bergholtz, P. Johannesson, P. Wohed (2010); Anchor modeling - Agile information modeling in evolving data environments, *in Data & Knowledge Engineering*, 69(12):1229-1253.

[14] R. Kimball and M. Ross (2013); *The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modeling*, Wiley, 3rd ed., 2013.

[15] M. Golfarelli and S. Rizzi (2009); *Data Warehouse Design - Modern Principles and Methodologies*, McGraw - Hill, 2009.

[16] Anchor Modeling Tool, available at: `http://www.anchormodeling.com/modeler`

[17] E. Malinowski and E. Zimanyi (2008); *Advanced Data Warehouse Design - From Conventional to Spatial and Temporal Applications*, Springer-Verlag Berlin Heidelberg, 2008.

[18] R. T. Snodgrass, I. Ahn (1986); Temporal Databases, *in IEEE Computer*, 19(9):35 - 42.

[19] C. Jensen, J. Clifford, R. Elmasri, S. K. Gadia, P. J. Hayes, S. Jajodia (1994); A Consensus Glossary of Temporal Database Concepts, *SIGMOD Record*, 23(1):52 - 64.

[20] H. Gregersen, J.S. Jensen (1999); Temporal Entity-Relationship models a survey, *IEEE Transactions on Knowledge and Data Engineering*, 11: 464-497.

[21] L. Ronnback, O. Regardt, M. Bergholtz, P. Johannesson, P. Wohed 92010); *From Anchor Model to Relational Database*, available at: `http://www.anchormodeling.com/wp-content/uploads/2010/09/AM-RDB.pdf`

[22] M. Golfarelli, J. Lechtenborger, S. Rizzi, G. Vossen (2006); Schema versioning in data warehouses: Enabling cross-version querying via schema augmentation, *in Data & Knowledge Engineering*, 59(2):435-459.

[23] B. Inmon (2004); *The Single Version of The Truth*, Business Intelligence Network (Powell Media LLC), available at: http://www.b-eye-network.com/view/282.

[24] R. Damhof (2008); *The next generation EDW*, available at: `http://prudenza.typepad.com/files/`

INTERNATIONAL JOURNAL OF COMPUTERS COMMUNICATIONS & CONTROL
ISSN 1841-9836, 12(2):183-200, April 2017.

Model of Network Topic Detection Based on Web Usage Behaviour Mode Analysis and Mining Technology

M. Chen

Mo Chen
1. Business College of Beijing Union University
A3, Yanjingdongli, Chaoyang district, Beijing, 100025, P.R. China.
mo.chen@buu.edu.cn
2. School of Information, Renmin University of China.
No.59, Zhongguancun Street, Haidian District, Beijing, 100872, P.R. China.
chenmoky@sohu.com

Abstract: This research has caught researchers' wide attention for detecting network topic exactly with the arrival of big data era characterized by semi-structured or unstructured text. This paper proposes a model of network topic detection based on web usage behaviour mode analysis and mining technology taking Web news as object of research. The author elaborates main function and method proposed in this model, which include the analysis module of Web news instance clicking mode, the analysis module of Web news instance retrieval mode, the analysis module of Web news instance seed and the analysis module of similar Web news instance supporting topics. Based on these functions and methods, the author elaborates main algorithm proposed in this model, which include the mining algorithm of Web news seed instances and the mining algorithm of similar Web news instances supporting topics. These functional algorithms have been applied in processing module of model, and focus on how to detect network topic efficiently from a large number of web usage behaviour towards to Web news instances, in order to explore a research method for network topic detection. The process of experimental analysis includes three steps, firstly, the author analyses the precision of topic detection under different method, secondly, the author completes the impact analysis of Web news topic detection quality from the number of Web news instances concerned and seed threshold, finally, the author completes the quality impact analysis of Web news instances mined supporting topic from the number of Web news instances concerned and probability threshold. The results of experimental analysis show the feasibility, validity and superiority of model design and play an important role in constructing topic-focused Web news corpus so as to provide a real-time data source for topic evolution tracking.
Keywords: web usage behaviour, network topic detection, clicking mode analysis, retrieval mode analysis.

1 Introduction

With the arrival of big data era, the field of information technology and Internet has developed a challenging stage so far. According to survey of TeckTarget that is a global leading professional IT network media [15] [2] [22], it has shown that the number of enterprises' data has broken through PB level with development of network, social media, business and other fields. Based on data existed and existing, people should think how to analyse complicated network data showing a tendency of explosive growth [4] [5], which have been concerned and are characterized by semi-structured or unstructured text, nevertheless, in whole process of cognizing network data, detecting topic exactly and effectively is the important and critical application direction.

In a mass of network data, the number of Web news released has reached EB level with events that continue to take place in social [6] [7], which shows the 4V features of big data, it is volume, variety, velocity and value [15] [9]. Based on these features above, Web news should reflect

Copyright © 2006-2017 by CCC Publications

high currency and reliability, on the basis of which, the topic contained in Web news should be condensed quickly and its path of evolution should also be tracked nearly in real time. However, how to detect network topic efficiently from a large number of web usage behaviour towards to Web news instances, it has become an urgent problem solved to construct a topic-focused Web news corpus so as to provide real-time data source for topic evolution tracking.

This paper proposes a model of network topic detection mainly containing four processing modules based on web usage behaviour mode analysis and mining technology taking Web news as object of research. The author elaborates function, method and technology on every processing module of the model in detail, which have been used or completed, and focuses on how to detect network topic efficiently from massive web usage behaviour towards to Web news instances. This process of research does key contribution for exploring a method for network topic detection, this experimental analysis results show the feasibility, validity and superiority of model design and implement.

2 Related works

In recent several years, some scholars have conducted some research about network topic detection method using different theory and technology. For example, Yang et all. survey research on the method of topic link detection based on improved information bottleneck theory [10], in this paper, a method of representing text is proposed, which can divide text into several sections of sub-topic features based on the regular pattern of semantic distribution and improve information bottleneck theory, then, the text represented by the attributes is utilized to do topic link detection, the experimental results have shown that this method has a fast convergent rate, and can improve the performance of topic link detection system. Suhara, Yoshihiko and others survey research on the method of information detection based on sentence-level topic [11], in this paper, the text sentence-level diversity features based on the probabilistic topic model is proposed, an information content classifier is also constructed combining features proposed, the experimental results show that this method outperforms the conventional methods. Pang, JB and others survey research on the method of unsupervised web topic detection using a ranked clustering-like pattern across similarity cascades [12], in this paper, a method using a clustering-like pattern across similarity cascades is investigated from the perspective of similarity diffusion, a topic-restricted similarity diffusion process is also proposed to identify real topic from a large number of candidates efficiently, the experimental results demonstrate that this approach outperforms the state-of-the-art methods on several public data sets, those works are related to author's research direction of network topic detection and application.

In recent several years, some scholars have also conducted certain research about method and technology of web usage behaviour analysis and mining. For example, Dziczkowski, Grzegorz and others survey research on the opinion mining approach for web user identification and clients' behaviour analysis [13], in this paper, an approach based on statistical analysis of natural language is proposed, three different methods are used for classifying opinions from clients' data, two new methods are introduced based on linguistic knowledge, in order to assign a mark dependent upon the client's emotions and opinions described in comments, the effect of experiments demonstrates that the system developed can carry out an evaluation and rating of opinions. Karakostas, Bill and others survey research on the MapReduce architecture for web site user behaviour monitoring in real time [14], in this paper, a MapReduce style architecture is proposed, where the processing of event series from the web users is performed by a number of cascading mappers, reducers, local to the event origin, the experimental results show that this architecture is capable to carry out time series analysis in real time for very large web data sets based on the actual events instead of resorting to sampling or other extrapolation techniques. Zhang,

YH and others survey research on the new replacement algorithm of web search based on user behaviour [15], this paper analyses the search ranking based on the user behaviour investigated mass distribution of information on the website, then proposes a replacement algorithm for web search, the simulation experimental results show that this approach under the search algorithm can reduce the execution time of retrieval effectively, and the optimal parameter selection for this blocking organization can be discussed continuously, those works are related to author's research direction of web usage behaviour application for analysis and mining.

Based on the analysis of the related research on network topic detection method and web usage behaviour application technology, experts and scholars have studied on two directions, but the research of constructing a network topic detection model based on web usage behaviour mode analysis and mining technology taking Web news as an object of analysis according to its attention and usage trait is missing. Therefore, this paper proposes a model of network topic detection based on web usage behaviour mode analysis and mining technology mainly, in order to explore how to detect network topic accurately.

3 Problem definition and notations

With rapid development of information and network technology, there are many types of network information, such as short text of micro blog, short, moderate or long text of Web news, long text of document and so on, while the biggest difference is structure of text content among them. This paper selects Web news as the object of research in view of ensuring high adaptability that the model of network topic detection based on web usage behaviour mode analysis and mining technology should have, in order to achieve the ideal effect of topic detection in the aspects of analysis precision and so on, this research provides scientific method for constructing and validating model of network topic detection.

3.1 Usage feature analysis

Users can search and browse Web news from different dimensions, granularities and frequencies, which have been extracted and analyzed, these processes have been elaborated in previous journal article published by authors [16], [17], [18], [19]. In the process of searching and browsing Web news, the user usage behaviour can be recorded, which not only explains Web news features used by users, but also contains the concerning topics hidden in Web news instances. Therefore, based on the analysis of Web news usage features, it is conducive to discover knowledge hidden in massive Web news, detect topic that the users are concerned about, track a series of events occurring in topic, and comb out process of event evolution.

From the perspective of global usage, every Web news instance concerned by users can be seen as a node in the range of Web news websites with authority, and the node set of some relevant instances supporting social events can be considered as a topic, each topic can also trigger a series of events, therefore, when users directly conceren a series of topics reflected by the social event, not only browse multi Web news instances content that support the topic, but also browse a series of events triggered by the topic [21]. From the perspective of local usage, when users search for Web news, in addition to input keywords that are related to the social events reported by Web news, but also input semantic keywords that may appear in Web news title or content, core event, core event occurring time, core event occurring location, subjective or objective object of the core event, and relation event information triggered by the core event [21]. Therefore, in the process of topic detection towards to Web news, if the usage features of Web news can be considered, then it can be mined for social events reported by Web news, topics

concerned by social events, Web news instances supporting topics, which will provide a Web news topic corpus with high quality for Web news topic evolution analysis.

3.2 Topic detection norm analysis

Based on the analysis towards to the usage features of Web news, if logical relationship need to be mined from behaviour data among the social events reported by Web News, the topics concerned in social events, Web news instances supporting topics, then the data and norms adopted should be specified in the process of topic detection [22], [23], which include user behaviour records, Web news clicking frequency based on S-U, Web news clicking frequency corresponding to URL, Web news clicking rate corresponding to URL and so on.

In the application process of Web news topic detection for social events, users can input the interesting keywords searched, when clicking submit request, Web news page will show multitudinous title, releasing time, releasing source and content link of Web news instances, and when users click on the contents link of the Web news instance, the application platform will record user names, the search information submitted, the behaviour clicking on Web news instances, the usage time and other data items, in which the search information submitted is expressed in English, but processed in this paper by the way of Chinese.

Based on the Web news usage behavior above, (s, u) of $S - U$ information can express retrieval keywords and Web news instances URL contained in behavior synchronously, $fq(s, u)$ can express Web news instances clicking frequency based on $S - U$, which explains the number of (s, u) appearance in a certain time period, $fq_i(u)$ can express Web news clicking frequency corresponding to URL, which explains the number of the Web news instance appearance in the i particle size of a certain time period, $fq(u)$ can express Web news instance clicking frequency in a certain time period as shown in formula 1, $rt_i(u)$ can express Web news clicking rate corresponding to URL as shown in formula 2.

$$fq(u) = fq_1(u) + \ldots + fq_i(u) + \ldots + fq_n(u) \tag{1}$$

$$rt_i(u) = \frac{fq_i(u)}{fq(u)} \tag{2}$$

Based on the Web news usage behavior above, it can be converted into a graph $G = (S, U, E)$, in which S can express the set of retrieval keywords submitted, U can express the set of Web news instances URL clicked, E can express the set of edges between S and U, the edge (s, u) can express behavior of clicking the Web news instance after submitting search request for users, whose weight value is $fq(s, u)$ corresponding to it.

$S(u)$ can express the set of S, which directly connect with u in G, $U(s)$ can express the set of U, which directly connect with s in G, $D(s)$ can express the degree of node in search request, which is the number of Web news instance nodes that is connected to retrieval requirement, D(u) can express the degree of the Web news instance node, which is the number of retrieval requirement nodes that is connected to the Web news instance as shown in figure 1, in which the S set is expressed in English, but processed in this paper by the way of Chinese.

3.3 Problem Notations

In this section, the author provides notations used in model and algorithms based on practical value and application direction of Web news topic detection. Let $NewsSet$ be a set of Web news instances, which is a data source using Web news for user and contains a large number of instances in Web news websites with authority. Let $UserBehavior$ be a set of records using Web news

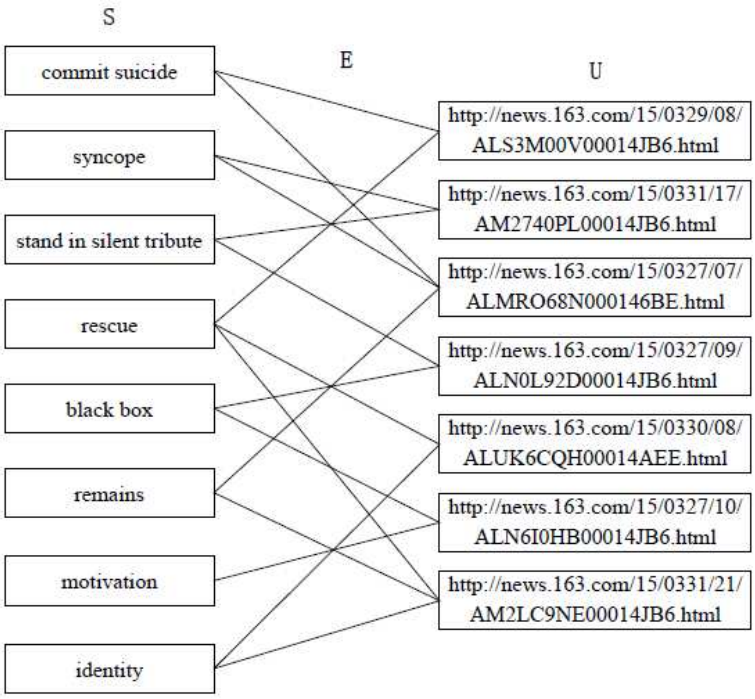


Figure 1: The behavior transition diagram based on Web news usage for users

behavior for users, which is also a data source of Web news topic detection, the specific notations are as follows.

Definition 1. Given a set of NewsSet, it can denote using $NewsSet = \{ns_1, \ldots, ns_i, \ldots, ns_k\}$, the range of i is between one and $k.ns_i.url$ stores address of the Web news instance, $ns_i.title$ stores title of the Web news instance, $ns_i.pubtime$ stores releasing time of the Web news instance, $ns_i.pubsource$ stores releasing source of the Web news instance, $ns_i.content$ stores content of the Web news instance, $ns_i.keyword$ stores keywords of the Web news instance, the extraction process of these information has been elaborated in previous articles published by authors [16], [17].

Definition 2. Given a set of UserBehavior, it can denote using $UserBehavior = \{ub_1, \ldots, ub_i, \ldots, ub_n\}$, the range of i is between one and $n.ub_i.username$ stores user names using Web news, $ub_i.searchword$ stores keyword retrieving Web news, $ub_i.url$ stores URL of the Web news instances clicked by users, $ub_i.systemtime$ stores system time using Web news for users.

Definition 3. Based on the definition and notations above, the problem that the Web news topic detection model needs to solve is to detect topic set contained in massive Web news instances from massive Web news usage behavior, and mine set of Web news instances that can support relevant topic, this result can denote using $TopicURL = \{tu_1, \ldots, tu_i, \ldots, tu_k\}$, the range of i is between one and $k.tu_i =< Topic, Topicurl >$, $tu_i.Topic$ can express topic description detected, $tu_i.Topicurl$ can express the set of Web news instances URL mined, which can support relevant topic detected.

4 The design of network topic detection model

In the era background of big data development, it has become an important research direction to detect network topic exactly in Web text mining field through the process of defining detection targets, extracting valuable network information, analysing web user usage behaviour, mining potential topics and applying topics detected and so on.

Based on this process, the model of network topic detection based on web usage behaviour mode analysis and mining technology taking Web news as object of research is divided into four modules, which include the analysis module of Web news instance clicking mode, the analysis module of Web news instance retrieval mode, the analysis module of Web news instance seed and the analysis module of similar Web news instance supporting topics as showed in figure 2.

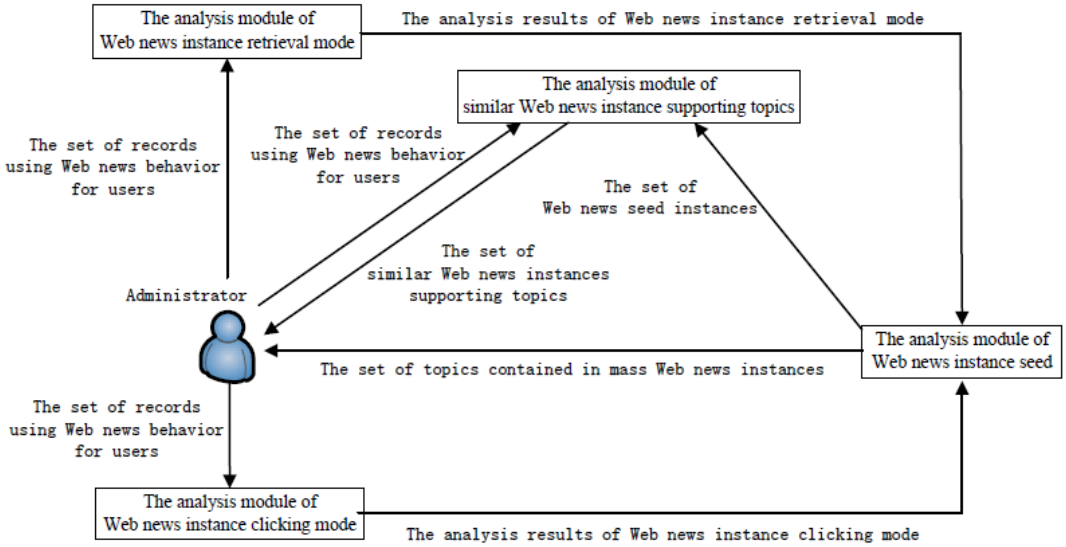


Figure 2: The model of network topic detection

4.1 The analysis module of Web news instance clicking mode

The inputting content of this module is set of records using Web news behaviour for users, the outputting content of this module is analysis results of Web news instance clicking mode, the main function of this module is to analyse the outbreak and concern mode of Web news instances according to records using Web news behaviour for users, and infer clicking mode of Web news instance.

Based on analysis of Web news usage behaviour, in a certain period of time, the Web news instance mays show a clear outbreak mode regarded as a kind of sensor for social event topic aggregation in records of user's Web news usage behaviour, which indicates whether the Web news instance is able to describe relevant topic reflected in social events. Therefore, in this module, firstly, administrator should use the outbreak mode of the Web news instance to measure whether it is able to describe corresponding topic reflected in social events, whose value is expressed with BR(u) that is a conjecture to sharpness of clicking rate changes, and its range is between 0 and 1 as shown in formula 3, this formula cites characteristic of entropy.

$$BR(u) = 1 - (-(rt_1(u)log_n rt_1(u) + \ldots + rt_i(u)log_n rt_i(u) + \ldots + rt_n(u)log_n rt_n(u))) \quad (3)$$

In the above formula, n represents the number of continuous granularity components clicked for Web news instances, the granularity can be set at hour or day for unexpected events in society, while the granularity can be set at week or month for normal events in society, if the fluctuations of clicking rate is not strong in the granularity setting for the Web news instance, then $BR(u)$ value is smaller, if the fluctuations of clicking rate is strong in the granularity setting for the Web news instance, then $BR(u)$ value is bigger.

Although using formula 3 can explain outbreak mode of the Web news instance, but its value cannot be absolutely certain that the key information of the Web news instance can describe topic in social events, because the $BR(u)$ value may be one, while the Web news instance is only once clicked by users. Therefore, in this module, secondly, administrator should use the concern mode of the Web news instance to measure whether it is able to describe corresponding topic reflected in social events again, whose value is expressed with CR(u), and its range is between 0 and 1 as shown in formula 4.

$$CR(u) = \frac{log(fq(u)) - Min_{u_i \in U}(log(fq(u_i)))}{Max_{u_i \in U}(log(fq(u_i))) - Min_{u_i \in U}(log(fq(u_i)))} \tag{4}$$

Because the process clicked has a characteristic of power law distribution for Web news instances, so in the above formula, the clicking frequency of the Web news instance is transformed to logarithm, if the Web news instance can be concerned by more users, then $CR(u)$ value is larger, whereas, $CR(u)$ value is smaller. Based on the above measurement of outbreak and concern mode for Web news instances, in this module, finally, administrator should use formula 5 to infer results of the Web news instance clicking mode.

$$ClickMode(u) = BR(u)CR(u) \tag{5}$$

4.2 The analysis module of Web news instance retrieval mode

The inputting content of this module is set of records using Web news behaviour for users, the outputting content of this module is analysis results of Web news instance retrieval mode, the main function of this module is to analyse the degree distribution and similar mode of Web news instances according to records using Web news behaviour for users, and infer the retrieval mode of Web news instance.

Based on analysis of behaviour diagram G using Web news for users, it can be found that the degree of Web news instances shows a characteristic of power law distribution, so in this module, firstly, administrator should use the degree distribution mode of Web news instance to measure whether it is able to describe corresponding topic reflected in social events and express it using $DR(u)$ as shown in formula 6.

$$DR(u) = \frac{log(d(u)) - Min_{u_i \in U}(log(d(u_i)))}{Max_{u_i \in U}(log(d(u_i))) - Min_{u_i \in U}(log(d(u_i)))} \tag{6}$$

Although using formula 6 can explain degree distribution mode of the Web news instance, but its value cannot be absolutely certain that the key information of the Web news instance can describe topic in social event, because the reason of generating Web news instance degree is that users search it using keywords. Therefore, in this module, secondly, administrator should use the similar mode of Web news instance to measure whether it is able to describe corresponding topic reflected in social events again, whose value is expressed with $SS(u)$, in order to solve problem of existing sparse records in user clicking behaviour as shown in formula 7.

$$SS(u) = \frac{2}{n(n+1)} Sum(\frac{Sum(s_{ik}(u)s_{jk}(u))_{k \in dataitem}}{\sqrt{Sum((s_{ik}(u))^2)_{k \in dataitem}}\sqrt{Sum((s_{jk}(u))^2)_{k \in dataitem}}})_{i<=j}^{n} \quad (7)$$

Based on the above measurement of degree distribution and similar mode for Web news instances, in this module, finally, administrator should use formula 8 to infer results of the Web news instance retrieval mode.

$$SearchMode(u) = DR(u)SS(u) \quad (8)$$

4.3 The analysis module of Web news instance seed

The inputting content of this module is analysis results of Web news instance clicking and retrieval mode, the outputting content of this module is sets of topic contained in massive Web news instances and Web news seed instances, the main function of this module is to infer set of Web news seed instances according to analysis results of Web news instance clicking and retrieval mode, and describe corresponding topic referring to Web news key information researched in previous job.

In this process, firstly, administrator should use formula 9 to mine set of Web news seed instances, its weight value is more than or equal to seed threshold, secondly, based on releasing time of Web news, the Web news seed instances should be sorted in set, finally, the corresponding topic should be described using key information of the Web news seed instance, in following experiment, the optimal value of seed threshold will be analysed.

$$SeedURL(u) = ClickMode(u)SearchMode(u) \quad (9)$$

4.4 The analysis module of similar Web news instance supporting topics

The inputting content of this module is set of Web news seed instances and records using Web news behaviour for users, the outputting content of this module is set of similar Web news instances supporting topics, the main function of this module continues to mine set of similar Web news instances with topic described according to sets of Web news seed instances, the corresponding topics described, the records using Web news behaviour for users that have been analysed above.

In this process, administrator should take Web news seed instances as core, and use the probability of first transfer from the Web news instance to itself as possibility that whether it is able to support topic described. If the Web news seed instance is set su, then the variable t_u indicates that whether the Web news instance is able to support topic described by su, the variable t_s indicates that whether the retrieval keyword is able to support topic described by su. If the key information of the Web news instance can support topic described by su, then $t_u = 1$, otherwise $t_u = 0$, if the retrieval keyword can support topic described by su, then $t_s = 1$, otherwise $t_s = 0$. For each Web news seed instance mined, in initial state, t_{su} is set one, then $P(t_{su} = 1) = 1$, and the probability is set zero for any other Web news instances supporting topic described by su, in this way, $P(t_s = 1)$ can be used to calculate the probability arriving su to itself, which is directly linked to the search keyword with it.

$$P(t_s = 1) = Sum(\psi_{su} P(t_u = 1))_{u:(s,u) \in E} \quad (10)$$

$$\psi_{su} = \frac{fq(s,u)}{Sum(fq(s,u_i))_{(s,u_i) \in E}} \quad (11)$$

In formula 11, ψ_{su} expresses transfer probability form the search keyword to the Web news instance, based on this probability, the value of $P(t_u = 1)$ can be calculated by using the following formula for all other Web news instances that are directly connected to it.

$$P(t_u = 1) = Sum(\psi_{us} P(t_s = 1))_{s:(s,u)\in E} \tag{12}$$

$$\psi_{us} = \frac{fq(s,u)}{Sum(fq(s_i,u))_{(s_i,u)\in E}} \tag{13}$$

In formula 13, ψ_{us} expresses transfer probability form the Web news instance to the search keyword, when $P(t_u = 1)$ is greater than or equal to probability threshold, then the Web news instance can be found to support topic described by su, in order to mine set of similar Web news instance that can support topics, in the following experiment, the optimal value of probability threshold will be analysed.

5 The design of network topic detection algorithm

Based on model design of network topic detection above, in this section, the author designs the mining algorithms of Web news seed instances and similar Web news instances supporting topics, in order to make sure that the topic detection has a high accuracy for Web news, in following experiments, the precision of algorithms will be analysed, the optimal value of parameters will be determined.

5.1 The mining algorithm of Web news seed instances

The key information of Web news have been expressed using Web news information extraction and analysis method researched in previous job [16], [17], [18], but what topics the users concern are still unknown in the face of massive Web news released based on social events. Therefore, this algorithm mainly uses set of records using Web news behaviour for users and results of Web news information extraction and analysis, and through analysing Web news instance clicking and retrieval mode to mine set of topic contained in massive Web news instances.

5.2 The precision analysis of topic detection under different method

This experiment compares precision of Web news topic detection under Web news instance clicking mode analysis method, Web news instance retrieval mode analysis method and the method proposed in this paper. As shown in figure 3, the precision represents quality of Web news topic detection using three methods, the red column expresses precision change situation of Web news topic detection using Web news instance clicking mode analysis method that is called DClickMode method in this chart, from its trend, it can be known that the quality of Web news topic detection is not high only through a single analysis for Web news instance clicking mode with increasing number of Web news instances concerned, although the precision has a certain improvement, but the maximum can only float on the 62.8% the blue column expresses precision change situation of Web news topic detection using Web news instance retrieval mode analysis method that is called DSearchMode method in this chart, from its trend, it can be known that the quality of Web news topic detection is not also high comparing with DClickMode method only through a single analysis for Web news instance retrieval mode with increasing number of Web news instances concerned, although the precision has also a certain improvement, but the maximum can also only float on the 63%, the green column expresses precision change situation of Web news topic detection using method proposed by this paper in this chart, from its trend, it can

be known that the quality of Web news topic detection has been significantly improved, because of integrating two analysis methods of Web news instance clicking and retrieval mode, while the quantity of Web news instances concerned is less, although the difference of precision is not big comparing with other two methods, the distance of precision is constantly widening among other two methods with increasing number of Web news instances concerned, the maximum can float on the 75.2%, this experiment shows that the quality of Web news topic detection using method proposed in this paper is higher than DClickMode and DSearchMode method.

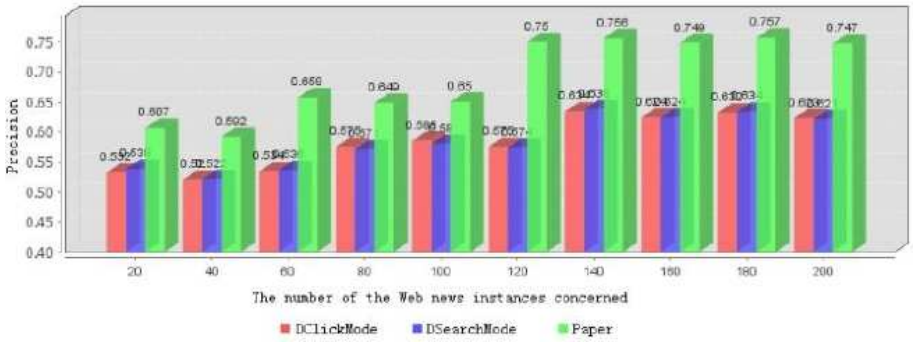


Figure 3: The quality of Web news topic detection under different methods

5.3 The impact analysis of Web news topic detection quality from the number of Web news instances concerned and seed threshold

Under the increasing number of Web news instances concerned, this experiment analyses precision change situation of Web news topic detection through adjusting seed threshold. As shown in figure 4, the precision represents quality of Web news topic detection with increasing number of Web news instances concerned in Y axis through adjusting seed threshold in X axis, when the threshold value is certain, from this graph, it can be known that the quality of Web news topic detection can increase slowly until a relatively stable trend with increasing number of Web news instances concerned, the main reason is that when the number of Web news instances concerned is less, the data and its relationship among them are relatively simple in the process of analysing clicking and retrieval mode, when the number of Web news instances concerned is increasing, the link relationship will exist among data in the process of analysing clicking and retrieval mode, which is conducive to topic detection, so that the precision of Web news topic detection is gradually increasing until more stable, when the number of Web news instances concerned is certain, from this graph, it can be known that the quality of Web news topic detection shows a trend of increasing firstly and then decreasing with increasing of threshold value, the main reason is that when the threshold value is less, a part of inaccurate or approximate accuracy Web news topic is likely to be detected as accurate Web news topic, when threshold value increases to a certain stable interval, only a small number of approximate accurate Web news topic are likely to be detected as accurate Web news topic, when threshold value is increased to a certain value, a part of accurate Web news topic may not be detected, this experiment shows that when the number of Web news instances concerned is 160, and seed threshold is 0.75, the quality of Web news topic detection can reach the highest value, which is close to 78.5%.

5.4 The mining algorithm of similar Web news instances supporting topics

Although the topics concerned by users have been detected using the mining algorithm of Web news seed instances, but what Web news instances supporting these topics the users concern

Algorithm 1 MiningSeedTopic

```
Input: UserBehavior, NewsSet, Threshold, InitialTime, T;
Output: TopicURL
MiningSeedTopic(UserBehavior ub, NewsSet ns, SystemData s);
BEGIN
UserRecord u[]=ExtractRecord(ub);
double br,cr,cm,dr,ss,sm;
GroupUserRecord<u> gur[];
TopicURL tu=new TopicURL();
gur=GroupByURL(u);
for i do=0 to gur.size()-1
    br=CalculateBR(gur[i]);
    cr=CalculateCR(gur[i],u);
    cm=br*cr;
    dr=CalculateDR(gur[i],u);
    ss=CalculateSS(gur[i]);
    sm=dr*ss;
    if c thenm*sm>=s.getThreshold()
        tu.add(ns.getFT(gur[i].url),gur[i].url);
    end if
    if g thenetCurrentTime()-s.getInitialTime()>=s.getT()
        ReSort(tu);
        ReAdjust(s.getThreshold());
    end if
end for
END
```

are still unknown in addition to Web news seed instances. Therefore, this algorithm mainly uses sets of records using Web news behaviour for users and Web news seed instances analysed to mine set of similar Web news instance which can support topics.

6 The experimental analysis and results

In this section, the author carries out experimental analysis and shows experimental results in order to validate feasibility, validity and superiority of model proposed in this paper, in this process, the author adopts experimental environment towards to event of German A320 plane crash shown as follows. The processor is dual core, the memory is 32G, the language of computer programming design is Java, its version is Java SE Development Kit 8, the platform of experimental design and implementation is MyEclipse 2015, the platform of experimental data storage and management is Microsoft SQL Server 2016.

6.1 The quality impact analysis of Web news instances mined supporting topic from the number of Web news instances concerned and probability threshold

Under the increasing number of Web news instances concerned, this experiment analyses precision change situation of Web news instances mined supporting topic through adjusting probability threshold. As shown in figure 5, the precision represents quality of Web news instances

Algorithm 2 MiningSimilarTopicURL

```
Input: TopicURL, UserBehavior, Threshold, InitialTime, T.
Output: TopicURL
MiningSimilarURL(TopicURL tu, UserBehavior ub, SystemData s)
BEGIN
double pts,ptu;
for i do=0 to tu.size()-1
    SearchWord sw1,sw2;
    WebNewsURL wnu1,wnu2;
    sw1=IsExist(tu[i].get("Topicurl"),ub);
    while s dow1!=NULL
        for j do=0 to sw1.size()-1
            pts=Calculate(sw1.get(j).position,tu[i].get("Topicurl"),ub);
            if p thents>=s.getThreshold()
                sw2.add(sw1.get(j));
            end if
            ub.set(sw1.get(j).position,pts);
        end for
        for j do=0 to sw2.size()-1
            wnu1=IsExist(sw2.get(j).position,ub);
            if ( thenwnu1=IsEqual(wnu1,wnu2))!=NULL
                for k do=0 to wnu1.size()-1
                    ptu=Calculate(wnu1.get(k).position,sw2.get(j).position,ub);
                    if p thentu>=s.getThreshold()
                        wnu2.add(wnu1.get(k));
                    end if
                    ub.set(wnu1.get(k).position,ptu);
                end for
            end if
        end for
        sw1=IsExist(wnu2,sw2,ub);
    end while
    tu[i].set(wnu2);
end for
if g thenetCurrentTime()-s.getInitialTime()>=s.getT()
    ReSort(tu);
    ReAdjust(s.getThreshold());
end if
END
```

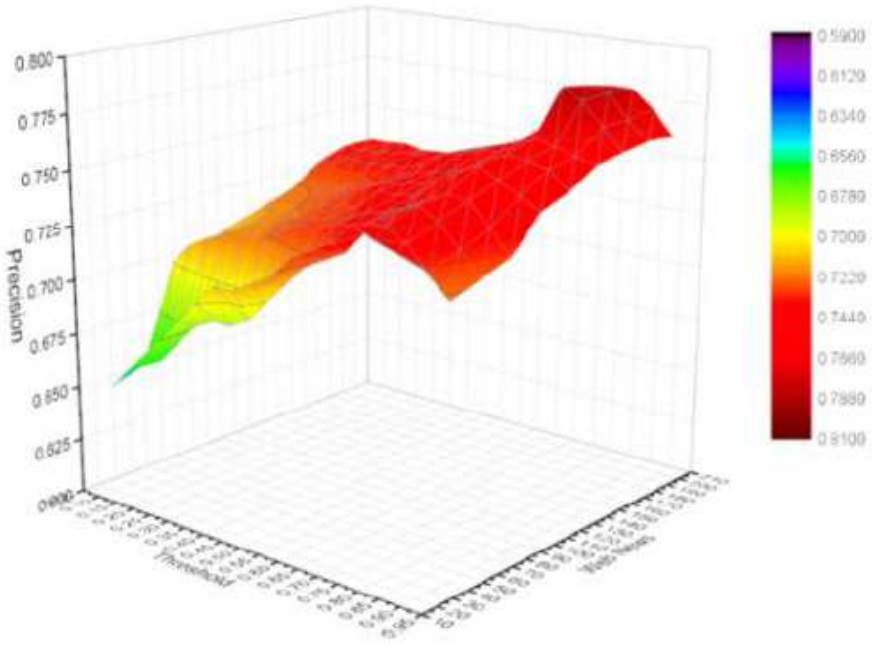


Figure 4: The changing trend of precision with number of Web news instances concerned and seed threshold

mined supporting topic with increasing number of Web news instances concerned in Y axis through adjusting probability threshold in X axis, when the threshold value is certain, from this graph, it can be known that the quality of Web news instances mined supporting topic can increase slowly until a relatively stable trend with increasing number of Web news instances concerned, the main reason is that when the number of Web news instances concerned is less, the data and its relationship among them are relatively simple in the process of analysing similar Web news instances supporting topic, when the number of Web news instances concerned is increasing, the link relationship will exist among data in the process of analysing similar Web news instances supporting topic, which is conducive to instance mine, so that the precision of Web news instances mined supporting topic is gradually increasing until more stable, when the number of Web news instances concerned is certain, from this graph, it can be known that the quality of Web news instances mined supporting topic shows a trend of increasing firstly and then decreasing with increasing of threshold value, the main reason is that when the threshold value is less, a part of inaccurate or approximate accuracy Web news instances are likely to be mined, when threshold value increases to a certain stable interval, only a small number of approximate accurate Web news instances are likely to be mined, when threshold value is increased to a certain value, a part of accurate Web news instances may not be mined, this experiment shows that when the number of Web news instances concerned is 140, and probability threshold is 0.7, the quality of Web news instances supporting topic can reach the highest value, which is close to 75.7%.

6.2 The process analysis of detecting Web news topic

The author illustrates effectiveness of Web news topic detection method implemented in this paper, As shown in figure 6, in this experimental webpage, firstly, users can choose the social event occurred that is German A320 plane crash, secondly, users can choose releasing time of Web news reporting the social event chose, thirdly, users can choose place, object or core event related

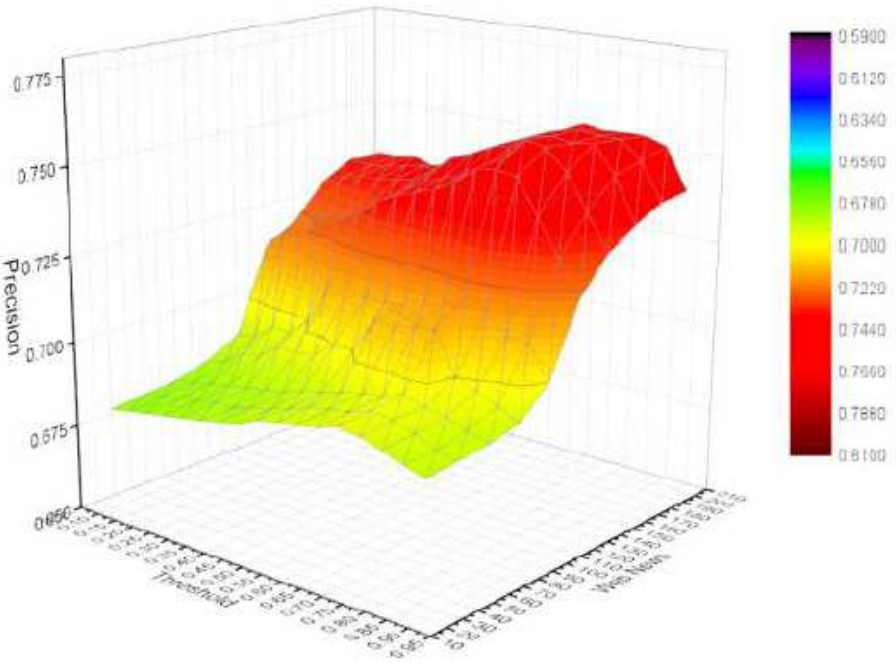


Figure 5: The changing trend of precision with number of Web news instances concerned and probability threshold

to the social event chose, whose selection range is imported from result of Web news extraction and analysis, fourthly, users can input keywords in textfield component again, finally, when users click Web news search button, the experimental platform will display webpage as shown in figure 7. In this webpage, the users can browse retrieval results of Web news instances according to retrieval condition chose or inputted that include title, releasing time, leading content and keyword of the Web news instances. When users are concerned about the Web news instance, and then click its title link, the experimental platform will display corresponding webpage that is linked source URL, at the same time, it will also record the current user name, the search keywords submitted, URL of clicking the Web news instance, usage time and other information in order to detect Web news topic.

As shown in figure 8, in this experimental webpage, firstly, users can choose the social event occurred that is German A320 plane crash, secondly, users can choose the number of topic detected that need to be displayed, finally, when users click submit button, the experimental platform will display top N topics description that have been detected, these topics are sorted in accordance with releasing time of corresponding Web news seed instance, in addition, it will also show set of Web news instances supporting every topic and category information of each topic, Web news instances are also sorted in accordance with its releasing time in set, when users are concerned about the Web news instance supporting topic, then click on its title link, the experimental platform will show corresponding Web news instance browsing webpage.

As shown in figure 9, in this experimental webpage, firstly, users can choose the social event occurred that is German A320 plane crash, secondly, users can choose the number of topic detected that need to be displayed, finally, when users click submit button, the experimental platform will show top N topics description detected by the way of time axis, these topics are sorted in accordance with releasing time of corresponding Web news seed instance, in addition, it will also show set of Web news instances supporting every topic, Web news instance are also sorted in accordance with its releasing time, when users are concerned about the Web news instance

supporting topic, then click on its title link, the experimental platform will show corresponding Web news instance browsing webpage.

Figure 6: The searching webpage of Web news

Figure 7: The clicking webpage of the Web news instance

Figure 8: The webpage of Web news topic detection

Conclusion

This paper completes a research on model of network topic detection based on web usage behaviour mode analysis and mining technology, which takes Web news as research object, takes web usage behaviour application technology as research core and executes process of defining detection targets, extracting valuable network information, analysing web user usage behaviour, mining potential topics and applying topics detected from point of innovation. This result is important and valuable for researchers in the same or related field. In the process of model research, design and implement, this paper proposes the mining algorithm of Web news seed instances and similar Web news instances supporting topics in order to eliminate shortcomings existing in previous traditional method.

Figure 9: The browsing webpage of Web news topics

The experimental analysis and results of model do key contributions for feasibility, validity and superiority of network topic detection request, improve efficiency of understanding network information for users, enhance availability of websites, build scientifically and improve service functions of websites, and improve business operational efficiency and clicking rate of websites. In a word, the process of research, design and implement model of network topic detection has certain practical application value, which establishes real and exact foundation of corpus for continuative research and application on Web text mining direction.

Acknowledgement

This work was supported by the National Natural Science Foundation of China under Grant Nos.71572015, the National Natural Science Foundation of China under Grant Nos.71271209, the Project of Philosophy and Social Sciences Planning in Beijing under Grant Nos.13JGC090.

Bibliography

[1] Zhang Ji, Li Hongzhou, Gao Qigang, Wang Hai, Luo Yonglong, Detecting anomalies from big network traffic data using an adaptive detection approach, *Information Sciences*, 6(3): 96-97.

[2] Pandey Suraj, Nepal Surya, Cloud Computing and Scientific Applications-Big Data, Scalable Analytics, and Beyond, *Future Generation Computer Systems*, 29(7): 1774-1775.

[3] Zhu Zhiguo, A novel method for discovering frequent changing patterns from historical web access data, *ICIC Express Letters*, 8(9): 2443-2445.

[4] Nasomyont, Tamrerk, A study on the relationship between search engine optimization factors and rank on google search result page, *Advanced Materials Research*, 3(4): 1462-1464.

[5] Guo Yi, Chen Hao, Microblog user ranking based on PageRank and Hadoop, *WIT Transactions on Information and Communication Technologies*, 49(1): 1083-1085.

[6] Zhang Hongli, Huang Shouming, Web Information Extraction Method Based on MapReduce, *Journal of Anhui Science and Technology University*, 27(2): 72-74.

[7] Li Wen, Zheng Bangxi, Deng Wu, Research on Web Information Extraction Model Based on XML and DOM Technologies, *Journal of Dalian Jiaotong University*, 34(3): 96-98.

[8] Zhang Yaming, Tang Chaosheng, Information propagation model based on the dynamics of complex networks in mircoblogging, *Journal of Computational Information Systems*, 10(1): 443-445.

[9] Wu Jiagao, Zhou Fankun, Zhang Xueying, Research of the Extraction Method of Event Properties Based on the Combining of HMM and Syntactic Analysis, *Journal of Nanjing Normal University(Natural Science Edition)*, 37(1): 30-32.

[10] Yang Yuzhen, Liu Peiyu, Fei Shaodong, Zhang Chenggong, A topic link detection method based on improved information bottleneck theory, *Zidonghua Xuebao/Acta Automatica Sinica*, 40(3): 471-479.

[11] Suhara, Yoshihiko, Toda, Hiroyuki, Nishioka, Shuichi, Susaki, Seiji, Automatically generated spam detection based on sentence-level topic information, *WWW 2013 Companion - Proceedings of the 22nd International Conference on World Wide Web*, 1157-1160.

[12] Pang Junbiao, Jia Fei, Zhang Chunjie, Zhang Chenggong, Unsupervised Web Topic Detection Using A Ranked Clustering-Like Pattern Across Similarity Cascades, *IEEE TRANSACTIONS ON MULTIMEDIA*, 17(6): 843-853.

[13] Dziczkowski, Grzegorz, Wegrzyn-Wolska, Katarzyna, Bougueroua, Lamine, An opinion mining approach for web user identification and clients' behaviour analysis, *IEEE Computer Society*, 79-84.

[14] Karakostas, Bill, Theodoulidis, Babis, A MapReduce architecture for web site user behaviour monitoring in real time, *DATA 2013 - Proceedings of the 2nd International Conference on Data Technologies and Applications*, 45-52.

[15] Zhang Yongheng, Feng Zhang, Fei You, A New Replacement Algorithm of Web Search Engine Cache based on User Behavior, *Applied Mathematics & Information Sciences*, 8(6): 3049-3054.

[16] Chen Mo, Yang Xiaoping, Research on Model of Network Information Extraction Based on Improved Topic-Focused Web Crawler Key Technology, *Tehnicki vjesnik/Technical Gazette*, 23(4): 49-54.

[17] Chen Xuegang, Research and realization of E-commerce monitor system based on focused web crawler, *Information Technology Journal*, 12(17): 4033-4035.

[18] Balla, Andoena, Real-time web crawler detection, *2011 18th International Conference on Telecommunications*, 428-430.

[19] Ahmadi-Abkenari, F, A clickstream-based web page significance ranking metric for web crawlers, *2011 5th Malaysian Conference in Software Engineering*, 223-225.

[20] Chen Mo, Yang Xiaoping, Liu Ting, A research on user behavior sequence analysis based on social networking service use-case model, *International Journal of u- and e- Service, Science and Technology*, 7(2): 1-4.

[21] Chen Mo, Yang Xiaoping, Sun Meng, Zhao Yun, Research on model of network information currency evaluation based on web semantic extraction method, *International Journal of Future Generation Communication and Networking*, 7(2): 103-105.

[22] Zhu Tao, Lin Yumin, Cheng Ji, Wang Xiaoling, Efficient diverse rank of hot-topics-discussion on social network, *Lecture Notes in Computer Science*, 8485(1): 522-524.

[23] Lu Ran, Xue Suzhi, Ren Yuanyuan, Zhu Zhenfang, A modified approach of hot topics found on micro-blog, *Lecture Notes in Electrical Engineering*, 269(1): 603-605.

INTERNATIONAL JOURNAL OF COMPUTERS COMMUNICATIONS & CONTROL
ISSN 1841-9836, 12(2):201-216, April 2017.

A Singleton Type-1 Fuzzy Logic Controller for On-Line Error Compensation During Robotic Welding

I. Davila-Rios, I. Lopez-Juarez, G.M. Mendez,
R. Osorio-Comparan, G. Lefranc, C. Cubillos

Ignacio Davila
COMIMSA - Posgrado Interinstitucional en Ciencia y Tecnologia (PICYT)
Ciencia y Tecnologia No 790 Col. Saltillo 400, C.P. 25290 Saltillo, Coahuila. Mexico.
idavila@comimsa.com

Ismael Lopez-Juarez*
Robotics and Advanced Manufacturing Group
Centro de Investigacion y de Estudios Avanzados del IPN (CINVESTAV)
Av. Ind. Metalurgica 1062, P. Ind. Saltillo-Ramos Arizpe, C.P. 25900 Ramos Arizpe, Coahuila, Mexico
*Corresponding author:
ismael.lopez@cinvestav.mx

Gerardo Maximiliano Mendez
Instituto Tecnologico de Nuevo Leon (ITNL)
Av. Eloy Cavazos 2001 Col. Tolteca, C.P. 67175 Guadalupe, Nuevo Leon, Mexico
gerardo.maximiliano.mendez@gmail.com

Roman Osorio-Comparan
Instituto de Investigaciones en Matematicas Aplicadas y Sistemas
Universidad Nacional Autonoma de Mexico (UNAM)
Circuito Escolar S/N, Ciudad Universitaria, Coyoacán, C.P. 04510, Mexico City
roman@unam.mx

Gaston Lefranc
Escuela de Ingenieria Electrica
Pontificia Universidad Catolica de Valparaiso
Avda Brasil 2950, Valparaíso, Chile 2430000
gaston.lefranc@pucv.cl

Claudio Cubillos
Escuela Ingenieria Informatica
Pontificia Universidad Catolica de Valparaiso
Avda Brasil 2950, Valparaíso, Chile 2430000
claudio.cubillos@pucv.cl

Abstract: During robot welding operations in the manufacturing industry there is a need to modify on-line the welding path due to a mismatch in the position of the components to be welded. These positioning errors are due to multiple factors such as ageing of the components in the conveyor system, clamp fixtures, disturbances, etc. Therefore, robot reprogramming is needed which requires a stop in the production line and consequently an increment in production costs. This article is an extension of [1][a] and presents an alternative solution to this problem that involves the use of structured lighting using a low-cost laser beam, a CMOS camera and a Gaussian singleton fuzzy logic controller. To validate the proposed control system, a robotic cell was designed using an industrial KUKA KR16 robot for welding metallic plates. The method was evaluated experimentally under lateral and vertical positioning errors.
Keywords: Gas Metal Arc Welding (GMAW), industrial robotics, artificial vision, robot path control, fuzzy logic.

[a]Reprinted (partial) and extended, with permission based on License Number 3947080516854 [2016] ©IEEE, from "Computers Communications and Control (ICCCC), 2016 6th International Conference on".

1 Introduction

The welding process establishes an electrical arc between a continuously fed electrode and the weld pool; which is protected by a gas administered externally, hence its name Gas Metal Arc Welding (GMAW) or commonly known as Metal Inert Gas (MIG). During the process, the molten electrode is transferred to the workpiece through the electric arc and serves as the filler metal (weld bead) which is deposited accurately by and automated mechanism, e.g. by

Copyright © 2006-2017 by CCC Publications

an industrial robot. During robot welding manufacturing operations, misalignments are likely to occur. There are several reasons; it can be due to disturbances, positioning sensing errors, ageing of welding fixtures and in general errors in positioning and conveying mechanisms. It is a common practice to make some robot reprogramming to solve the problem that requires the stop of the production line with an increment in production costs. There are currently commercial solutions that can be adapted to solve these common errors, however the costs can be very high, and in the order of the price of the industrial robot itself, which may not be affordable for some small companies. In this article, we present an alternative solution to this problem that involves the use of structured lighting employing a low-cost laser beam, a CMOS camera and a PC-based fuzzy controller.

1.1 Related work and objective

The work presented in this paper is an extended version of the originally published article in the IJCCC [1]. Similar work on robot control for seam tracking has been presented by Graf et al. [2] who developed a trajectory-based control for seam tracking by modelling the trajectory as a continuous curve in 3D. They showed good results; however, the technique has some drawbacks in cases where the solution does not exist for correcting the orientation of the seam location or if the current robot location is close to the last location. Santti et al. [3] have recently tested high performance processors to extract the dominant line from the segmented data of the trajectory. The system can achieve a line extraction speed of more than 1000 fps, which enables real-time visual seam tracking and robot control. Some approaches that only use an optical sensor have been reported. Liu et al. [4] have proposed to use optical filtering and the modification of the camera's exposure time to extract the geometrical profile of the seam. Some other approaches also have appeared in the literature that addresses neuro-fuzzy controllers that can be applied to robot positioning control [5].

The objective of the research presented in this paper is to correct the robot's end effector position online to eliminate the need of reprogramming the robot. In this manner, errors due to disturbances, positioning sensing errors, ageing of welding fixtures and in general errors in positioning and conveying mechanisms are eliminated. Our proposal is based on previous approaches using structure lighting (i.e. laser sensor) and image processing to quantify the misalignment and to react before the actual beam is formed by sending robot commands to reposition the arm robot so that initial offset cannot affect the next piece to be weld in a production line. The performance of the fuzzy controller is verified experimentally.

The paper is structured as follows. In section 2, the test bed is explained. The image processing for detecting the part misalignment is described in Section 3, whereas the design of the controller and its experimental results are explained in sections 4 and 5, respectively. Finally, conclusions, current and envisaged work are given at the end of the paper in Section 6.

2 Test bed

The following equipment composes the test bed: a KUKA robot arm manipulator, GMAW welding station, wire feeder that controls the wire supply to the torch, a PC-based data acquisition system and an inert gas tank as it is depicted in Figure 1

The test bed also includes a video camera and a laser sensor oppositely positioned to each other. By using this configuration, occlusions by the torch itself can be avoided. However, a perspective error occurs which is compensated by a homography matrix when using triangular singleton type-1 (T1) fuzzy logic controller (FLC) and it is directly compensated without the homography when a Gaussian singleton T1 FLC controller is used. With our proposal using the

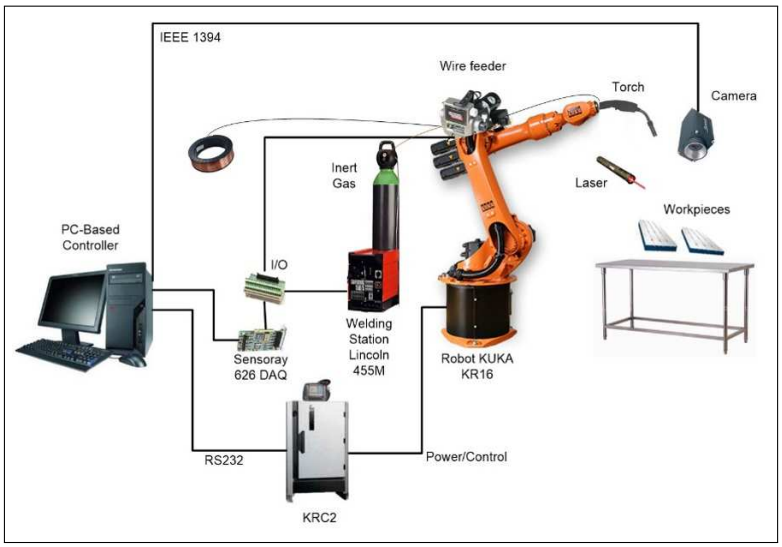


Figure 1: Welding robotised cell.

Gaussian singleton T1 FLC, the use of algorithms for perspective error correction and image processing of high computational costs are avoided.

The welding cell is controlled by a PC-based controller. The computer has an Intel Xeon @ 1.86 GHz processor with 3GB RAM that hosts the application interface and which also includes the following functions:

Serial communication. For on-line modification of the welding path. The communications are carried out at low-level between the PC and the KRC2 controller using the 3964 protocol.

Image processing. Filtering and segmentation are carried out using a CMOS Basler A602fc camera as input device with a spatial resolution of 656x490 pixels.

Voltage and current modification. The arch current and voltage is modified by DAQ Sensoray 626 using its I/O port. Different I/O signals are used to control the on/off of the welding station, the wire feeding system (including the open/close state of the gas valve), the laser sensor on/off condition, the welding travel speed and the distance from the tip of the torch to the workpiece.

The robot manipulator is used in slave mode. During the operations, a robot positioning program is run in the KRC2 controller that continuously search for motion commands from the PC controller in order to start an incremental motion of the robot arm. This program also controls the selection between tool and world coordinates, and the speed and motion step size during incremental motions. The positioning fuzzy controller resides in the PC controller sending the path modification commands to the robot controller during welding operations.

3 Image processing

During image processing some distortion can occur due to the alignment of the camera with respect to the image plane as it is shown in Figure 2.

A homography relates two images in perspective, where plane points from one scene are related to the second image. This relationship is valid if the scene is flat or the displacement small [6] as it is illustrated in Figure 3.

In this manner the homography helps to eliminate the perspective from one image and also

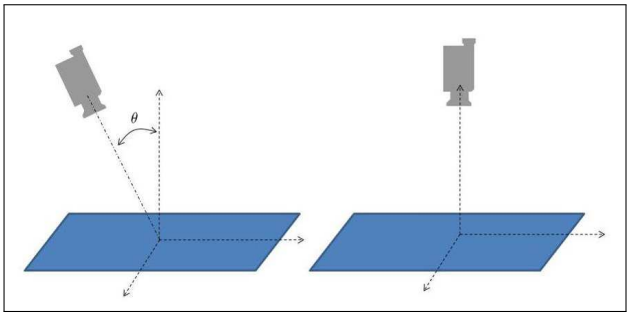


Figure 2: Camera alignment

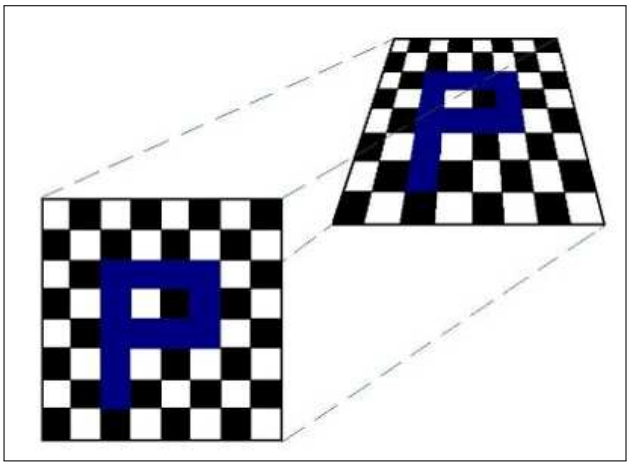


Figure 3: Image perspective

to facilitate the calibration. The homography map points (equation 1) from the work plane ($z = 0$) to the points in the image plane (equation 2). The homography components depend on the intrinsic camera parameters and its location in the space (extrinsic parameters), which are constants.

$$X = [X, Y, 1]^T \tag{1}$$

$$X' = [X', Y', 1]^T \tag{2}$$

The equation that represents the homography transformation is given by equation (3), where the homography matrix H can be obtained by using four points in the calibrated image and the image in perspective [6].

$$|X'| = |H||X| \tag{3}$$

In the test bed, typical errors occur when two metallic plates to be welded are misaligned. The first step to correct the situation is to measure this misalignment using image processing tools. Once the positional misalignment is quantified, the information is sent to the robot controller for compensation. The methodology consists of using a laser beam (50 mW with $\lambda = 656nm$) aimed to both plates as it is shown in Figure 4.

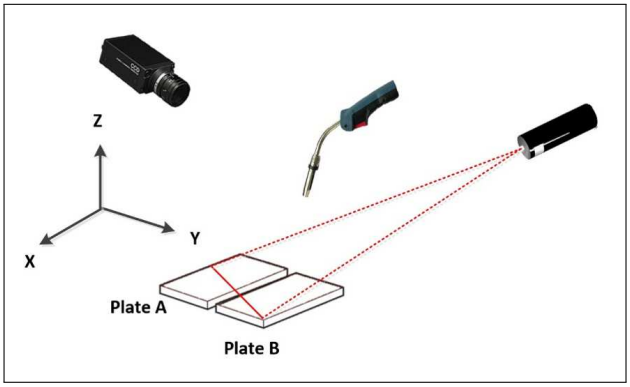


Figure 4: Welding methodology

When the laser beam is projected onto both plates, a discontinuous line appears and it is captured by the camera sensor as a reference signal. Misalignments can occur in two directions, either they can be lateral misalignments or height misalignments. In both cases, the misalignment is captured in the image. In order to accurately process this information, the image pre-processing is carried out first by smoothing the image applying a mean filter and max/min filtering to reduce noise.

After the pre-processing stage, the image is segmented in the region of interest (ROI) where the necessary information is found (pixel $I_{i,j}$ in Figure 5).

The main idea is to take an image of dimension KxL from the original I image with dimension IxJ, so that $I(K$x$L) < I(I$x$J)$. Having a reduced image in size is also useful to speed up the processing time. The segmented ROI is shown in Figure 6. The gap between Plate A and Plate B can be observed within the rectangle formed by the dashed lines.

Once the gap is determined, the next step is to determine $Discontinuity1$ and $Discontinuity2$ from the laser pattern and captured from each plate as showed in Figure 7.

During normal welding operation, the robot's torch should be located in the middle point between the plates as it is indicated (Robot) in Figure 8. In the case of any offset either lateral

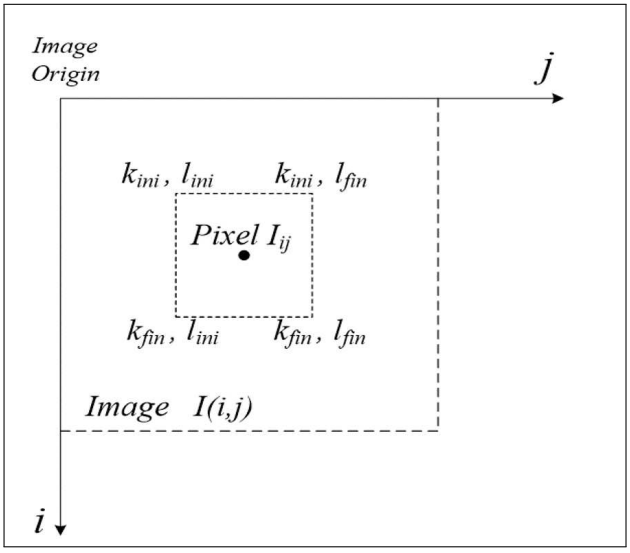


Figure 5: Image Segmentation

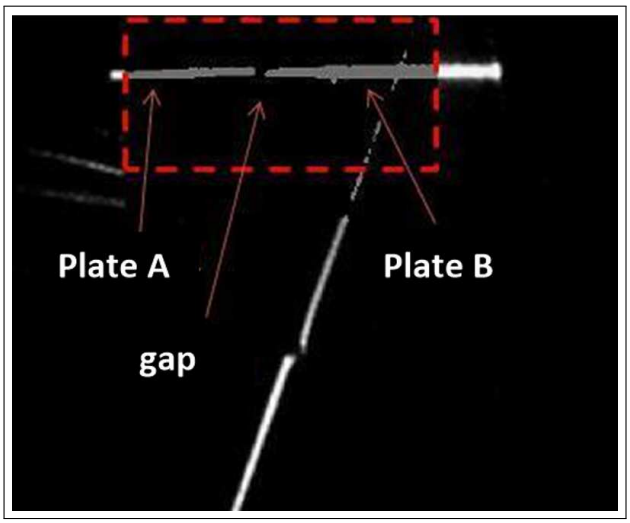


Figure 6: ROI with segmented image

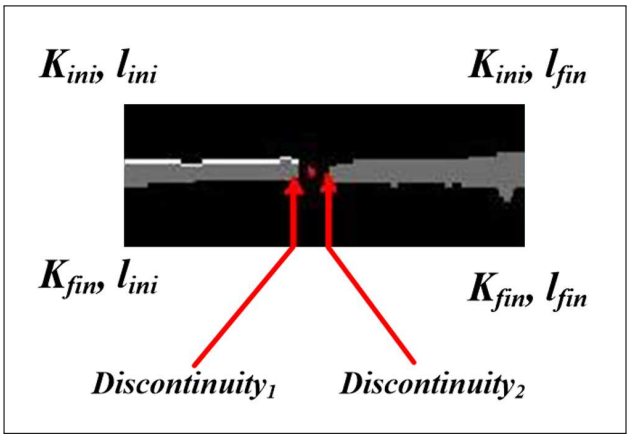


Figure 7: Discontinuity between plates

misalignment or misalignment in the Z axis direction, the point will move accordingly, hence requiring a repositioning strategy.

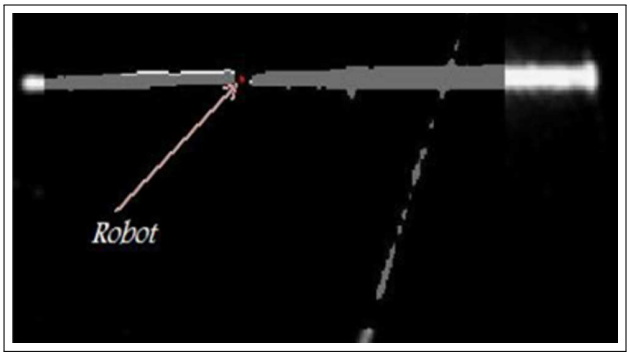


Figure 8: Point without offset position

In situations where the robot's torch is not in the correct position, the situation has to be assessed first by measuring the misalignment. A corrective motion will be a distance between points union (x, y) and robot (x, y) as depicted in Figure 9. The corrective motion has to be in any direction within 3D space volume, so that we proposed a Fuzzy algorithm to correct it.

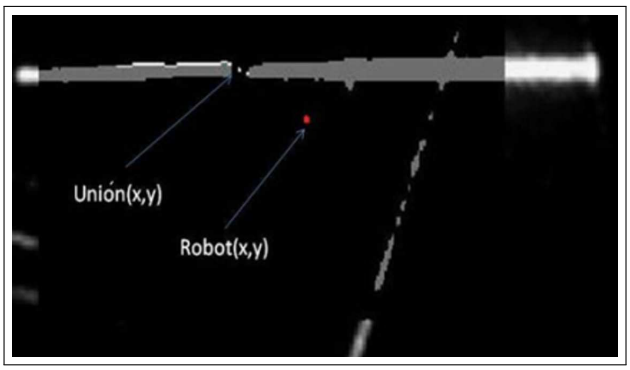


Figure 9: Torch offset position

4 Design of the controller

In our experiments two controllers were tested. The first controller is based on the Type-1 Fuzzy Logic Control which uses triangular and trapezoidal membership functions. The second controller uses Gaussian membership functions and is referred to as Singleton Fuzzy Logic Controller.

According to Mendel [7] membership functions characterize fuzzy sets. A type-1 (T1) fuzzy set, A which is in terms of a single variable $x \in X$, is a generalization of a crisp set. It is defined on a universe of discourse X and is characterized by membership function $\mu_A(x)$ that takes on values in the interval [0,1]. A membership function provides a measure of the degree of similarity of an element in X to the fuzzy set. Such a set may be represented as

$$A = \{(x, \mu_A(X))|\forall x \in X\} \tag{4}$$

$T-1$ membership function, $\mu_A(x)$ is constrained to be between 0 and 1 for all $x \in X$, a Gaussian membership function is specified by two parameters m, σ as follows:

$$gaussian(x : m, \sigma) = exp(-\frac{(x-m)^2}{\sigma^2}) \tag{5}$$

where m and σ denote the center and width of the function, respectively. The shape of the function can be controlled by adjusting the parameter σ. A small value of σ will generate "thin" membership function, while a big σ will lead to a "flat" membership function.

Fuzzy logic controllers (FLC's) are useful control schemes for plants having difficulties in deriving mathematical models or having performance limitations with conventional control schemes. Error e and change of error e' are the most used fuzzy input variables in most fuzzy control works, regardless of the complexity of controlled plants. Also, either control input ${}^{u}(PD-type)$ or incremental control input ${}^{\Delta u}(PI-type)$ is typically used as fuzzy output variable. T1 FLC's are both intuitive and numerical systems that map crisp inputs to a crisp output. Every FLC is associated with a set of rules with meaningful linguistic interpretations, obtained from either numerical data, or experts. Based on this kind of statement, actions are combined with rules in an antecedent-consequent format, and then aggregated according to approximate reasoning theory, to produce a nonlinear mapping from input space $X1 \times X2$ to the output space Y.

4.1 Input-output relationship

A fuzzy control in the general case is based on linguistic variables to handle imprecise and vague information. The information is embedded in fuzzy sets that are combined in rules to define actions to be taken as it is indicated in [8] and [9]. In our case, we need to define a set of input-output relationships in order to quantify the positional error of the robot arm with respect to the plate's welding position. In other words, the input information for the robot is the misalignment of the robot's end effector and the output is a robot position command to correct such misalignment.

The design of the tracking system is focused on the offset compensation in automatic production lines, where the parts to be welded are moved by pallets to other welding stations and where typical positioning errors are in the range of few millimetres. These errors are likely to occur between the current workpiece and the next due to several reasons as mentioned earlier. During these circumstances, the user would normally correct the robot path by reprogramming the robot which is time consuming. Our proposal is to carry out the correction on-line, using the same welding program. The idea is to modify incrementally the robot's offset path at the starting of the welding operation should a misalignment is detected so that not reprogramming is needed. In order to modify the path a deviation measurement has to be compared to the original path, so that the robot "knows" where to move to and for how much using linguistic variables such as "right", "left", "far left", "decrease slightly", "increase greatly", etc. The linguistic variables are translated into fuzzy sets with a membership function that considers a value in the range $[0, 1]$.

4.2 Fuzzy design

The steps to be considered in the design are normalisation, fuzzification, determination of fuzzy rules, defuzzyfication, and denormalisation. The first step is accomplished by using the information given from a sensing system, which in this case is obtained from the laser beam pattern as detected by the camera. This value is read and must be normalised to the range [0,1]. The next step is to fuzzify this information to assess its membership within the fuzzy sets. In our case the variable depends on the workpiece misalignment and we should find its membership function within the following fuzzy sets: far left, left, near left, near, near right, right, and far right.

Once the input data has been normalised and fuzzified, then this is used to build up the fuzzy rules what is meant to be the control antecedent. These rules are built depending on the requirement of the system and after its evaluation their membership function is determined in relation with the output set (consequent). In order to have useful singleton real values, the data set is defuzzified to obtain a new robot coordinate which is in turn sent to the robot controller to ultimately correct the welding trajectory.

If we consider the X axis to be the welding direction as depicted in Figure 4, it is clear that the misalignment would come only in the Y or Z axis direction. Considering this assumption, the control system is based on two input variables and two output variables. The operation range is defined in the interval $[-10, 10]$ mm in the Y axis and $[-5, 5]$ mm in the Z axis. The zero value is considered to be the reference value for the welding path. Figure 10 shows a set of input data in the Y and Z axes during the evaluation tests of the controller. These values comprise the whole set of values likely to be encountered during operations. For instance, if the algorithm detects a misalignment point (10, -5) it means that the workpiece had an offset of 10mm to the right and -5mm downwards that needs to be compensated.

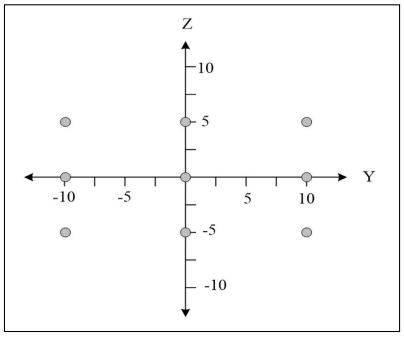


Figure 10: Robotic motion range for error compensation

The trapezoidal and triangular membership functions are used to design the input and output fuzzy sets as shown in the Figure 11.

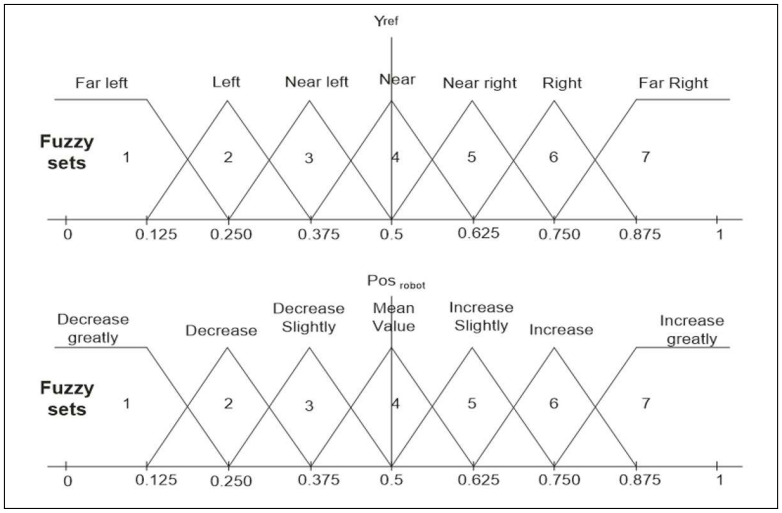


Figure 11: Fuzzy sets

After the degree of membership of the input values from the input fuzzy sets is defined, the antecedent is created using fuzzy rules as follows:

IF	Yref	IS	Far left	THEN	Decrease greatly	Posrobot
IF	Yref	IS	left	THEN	Decrease	Posrobot
IF	Yref	IS	Near left	THEN	Decrease slightly	Posrobot
Etc.	...	...	...	...	...	...

During the defuzzyfication stage the required non-fuzzy real values are obtained using the Centre of Area (CoA) method as a fuzzy conclusion "Y is A". This can be determined by equation 6 in the discrete domain.

$$y = \frac{\sum_i \mu_A(y_i) \times y_i}{\sum_i \mu_A(y_i)} \tag{6}$$

where $\mu_A(yi)$ is the membership function.

Having defuzzified the data, their values are denormalised obtaining a real value which is used as a new robot coordinate. The fuzzy robot controller will also consider other errors related to other variables such as ageing of the positioning mechanisms or disturbances. The correction is on-line, during the welding process avoiding the need of stopping the production line. The operating working range is $\pm 10mm$ in Y axis and $\pm 5mm$ in Z axis. The fuzzy controller was developed in C++ using the Visual Studio compiler.

4.3 Performance assessment

In order to assess the statistical performance of the controller a set of experiments were carried out. The robot path tracking ability during the operation range was evaluated $\pm 10mm$ in Y axis and $\pm 5mm$ in Z axis. The evaluation helps to analyse its behaviour against any variation of the experimental factors. The factors are considered either combined or in its individual form so that the interaction can be identified as indicated in [7]. Care was taken to consider two important aspects during the experimental design that are replicate and aleatorisation. The use of replicates is very important to determine the experimental error. The aleatorisation allows confirming that the random probability variables refer to independent probability distributions. Considering the above assumption, the input variables to the system are the position values that are sent to the robot in the Y and Z axis direction ($DistY$, $DistZ$) and as output, the real distance observed in both axes (DistY real and DistZ real).

Experimental design considers variables with two or three levels and k factors referred as to $2K$ or $3K$, respectively. In our experiments we decided to use 32 with 2 replicates. The decision to use three levels was based on the interest of using the central point within the robot's range motion in both, the Y axis and the Z axis.

5 Results

5.1 Results with the T1 FLC

The experimental procedure can be observed in Figure 12. The experimental set up and the coordinate frame are shown. The offset value (DistY, DistZ) that the fuzzy controller has to compensate is showed in the dashed rectangle. The output value is considered as the real value measured along the main axes (DistY real, DistZ real). It is important to note that the welding seam is applied along the X axis.

Table 1 contains the results from 18 experimental runs. The first 9 corresponds to the first replicate and runs 10 to 18 correspond to the second replicate. Figure 13 shows results obtained during the correction of lateral misalignment and the measured absolute error. Similarly, Figure 14 shows the obtained results during the correction in the vertical distance and its absolute

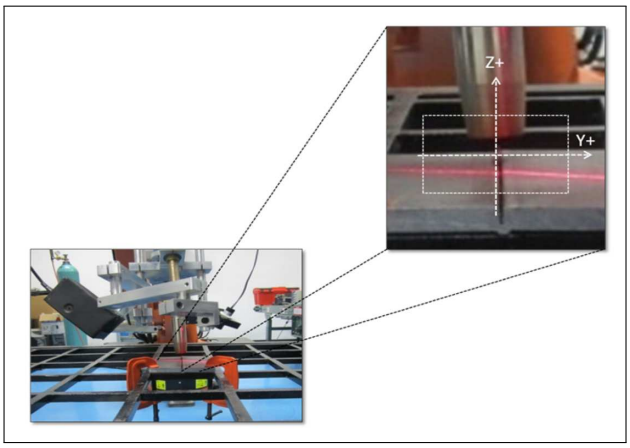


Figure 12: Experimental set up

error. The measured error demonstrates that the robot fuzzy controller compensates the misalignment in all cases. The maximum observed error is $1.6mm$ in Y axis, which is considered to be appropriate for practical purposes in industrial manufacturing processes.

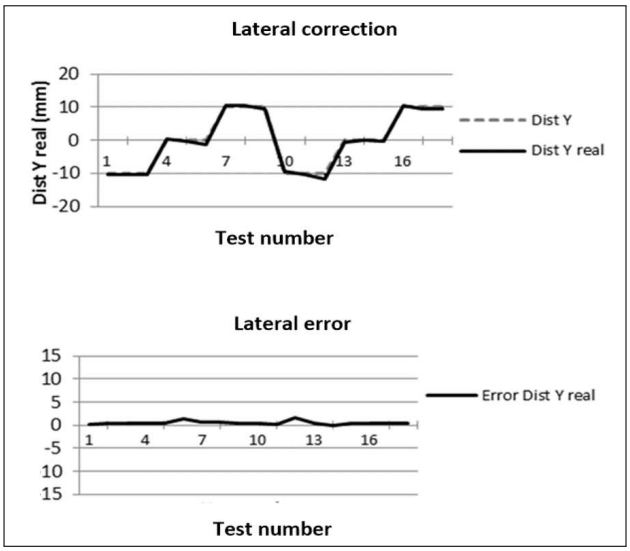


Figure 13: Lateral misalignment and absolute error

In comparison with the triangular T1 FLC, the Gaussian T1 FLC uses a bigger area of exploration. It was decided to duplicate the area 40mm x 20mm. The lateral misalignment range is +/− 20 mm in the Y axis whereas the vertical misalignment is +/− 10 mm in the Z axis. This can be observed in Figure 15.

The Gaussian T1 FLC uses more information as it considers mid points between the initial and final points. Every point within the exploration area forms a fuzzy rule. If we consider that the points are spaced 2mm then we will have 21 points in Y axis by 11 in the Z axis making a total of 231 points or fuzzy sets for the Singleton FLC. The Gaussian type of fuzzy sets are illustrated in Figure 16.

The Gaussian controller has the advantage of not requiring the homography since the input information is given directly from the camera in pixels. For the reference point (0,0) which is supposed to be the point where the welding torch should be aligned with is located in the point

Table 1: Experimental results and error evaluation (mm)

DistY	DistY Real	Error Y	DistZ	DistZ Real	Error Z	AbsY error	AbsZ error
-10	-10,2	0.2	-5	-5	0	0.2	0
-10	-10.4	0.4	0	-0.4	0.4	0.4	0.4
-10	-10.4	0.4	5	4.4	0.6	0.4	0.6
0	0.47	-0.47	-5	-5.26	0.26	0.47	0.26
0	-0.43	0.43	0	-0.42	0.42	0.43	0.42
0	-1.3	1.3	5	5.28	-0.28	1.3	0.28
10	10.6	-0.6	-5	-5.2	0.2	0.6	0.2
10	10.6	-0.6	0	0.6	-0.6	0.6	0.6
10	9.6	0.4	5	5.2	-0.2	0.4	0.2
-10	-9.5	-0.5	-5	-4.95	-0.05	0.5	0.05
-10	-10.2	0.2	0	-0.2	0.2	0.2	0.2
-10	-11.6	1.6	5	4.9	0.1	1.6	0.1
0	-0.53	0.53	-5	-5.3	0.3	0.56	0.3
0	0	0	0	-0.1	0.1	0	0.1
0	-0.33	0.33	5	5.27	-0.27	0.33	0.27
10	10.5	-0.5	-5	-5	0	0.5	0
10	9.6	0.4	0	-0.4	0.4	0.4	0.4
10	9.6	0.4	5	5	0	0.4	0

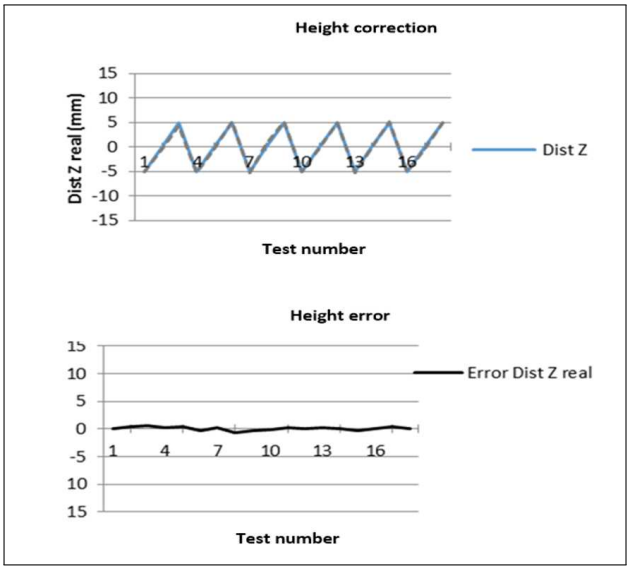


Figure 14: Vertical misalignment and absolute error

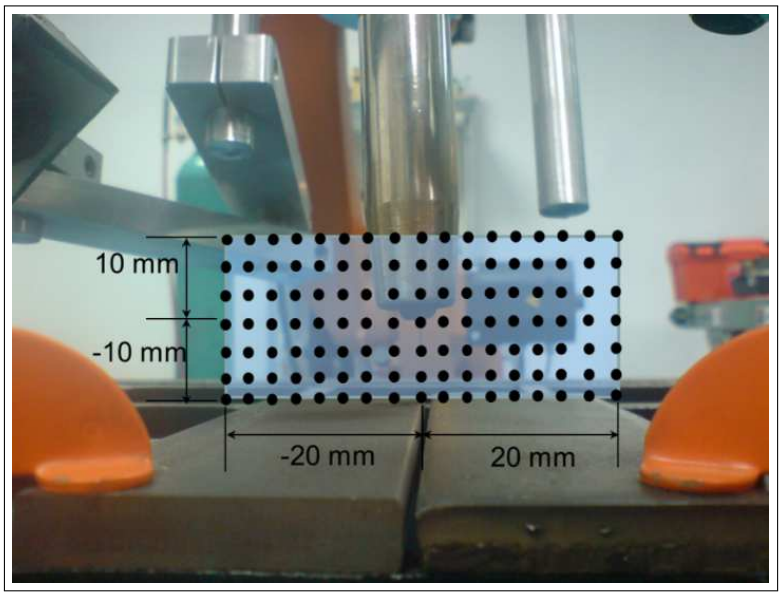


Figure 15: Exploration area in the Y and Z axis

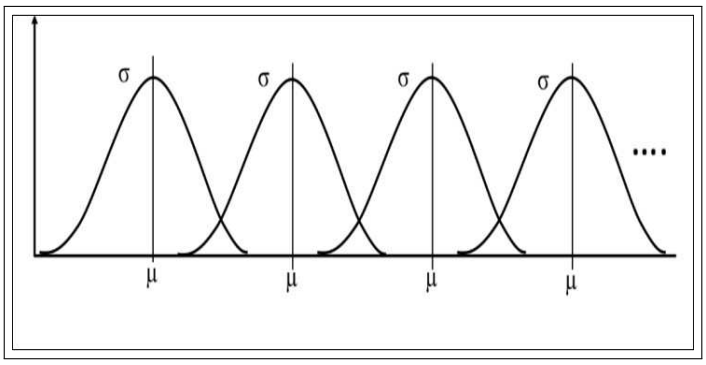


Figure 16: Gaussian type fuzzy sets

Table 2: Results from the Singleton FLC

No.	Y axis	Z axis	$PosRobot_Y$	$PosRobot_Z$	Error Y axis	Error Z axis
1	-20	10	19.59255	-9.99768	0.40745	0.00232
2	-20	0	19.98974	0.87296	0.01026	0.87296
3	-20	-10	19.8391	10	0.1609	0
4	0	10	0.02595	-10	0.02595	0
5	0	0	-0.02274	0.00005	0.02274	0.00005
6	0	-10	0.02708	9.99652	0.02708	0.00348
7	20	10	-19.90655	-10	0.09345	0
8	20	0	-19.96467	0.0156	0.03533	0.0156
9	20	-10	-19.99986	10	0.00014	0
10	-19	9	19.79273	-9.99535	0.79273	0.99535
11	-19	10	17.99499	-10.04568	1.00501	0.04568
12	-17	10	18.09578	-10	1.09578	0
13	-10	5	9.16485	-5.54219	0.83515	0.54219
14	-10	0	10.45749	-0.00265	0.45749	0.00265
15	-10	-5	10.87992	4.61136	0.87992	0.38864
16	0	5	0.197392	-5.95809	0.197392	0.95809
17	0	0	-0.0322	0.05833	0.0322	0.05833
18	0	-5	0.07501	5.80521	0.07501	0.80521
19	10	5	-10.03783	-5.96332	0.03783	0.96332
20	10	0	-9.96102	0	0.03898	0
21	10	-5	-9.15139	4.04127	0.84861	0.95873

(3.34598, -3.07248) pixels and so on with the other 230 points located within the total area. Once all points are related within the image, then the next step is to form the fuzzy sets by using the Gaussian function given by equation (5).

Since the information comes from the process and does not require normalisation, the number of fuzzy sets are 231 and its membership will be given by the mean and standard deviation from each point. In order to generate the fuzzy sets it is required to carry out at least 10 tests in each point, so the total number of tests are 2,310 points to generate the fuzzy sets.

To validate the performance of the Gaussian T1 FLC several tests were carried out within the exploration area. The position error of the torch in the Y and Z axis are the input variables (given in pixels), while the output variables are the robot's end effector coordinates given by $PosRobot_Y$ and $PosRobot_Z$. The Table 2 shows the results obtained after 21 trials. It can be observed that errors are lower compared to the case of the triangular T1 FLC and having a maximum value of approximately 1mm.

Figures 17 and 18 show the corresponding error graphs where it is clear that the misalignment of the plates was corrected and in all cases the repositioning of the arm had an error lower than 1mm, which is better that in the case of the triangular T1 FLC [1].

Conclusions

This article presented an alternative solution to this problem that involves the use of structured lighting using a low-cost laser beam, a CMOS camera and a singleton type-1 fuzzy logic controller. To validate the proposed control system, a robotic cell was designed using an industrial KUKA KR16 robot for welding metallic plates. The method was evaluated experimentally

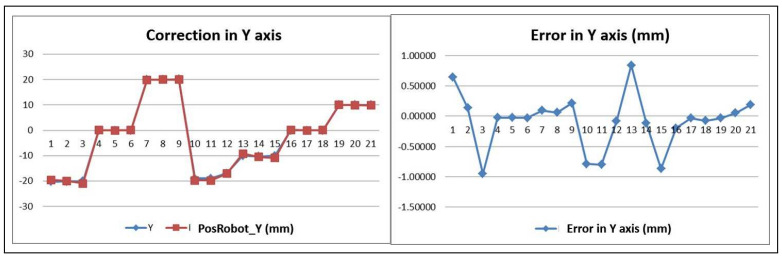


Figure 17: Relationship between offset error and correction in the Y axis

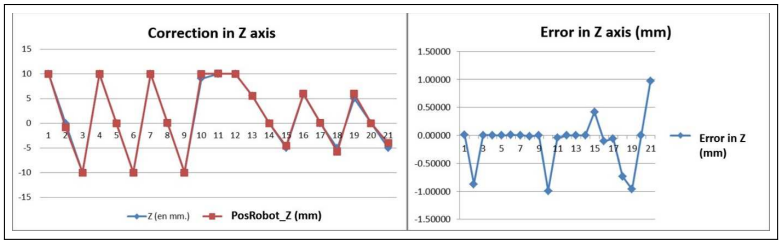


Figure 18: Relationship between offset error and correction in the Z axis

under lateral and vertical positioning errors. The methodology presented is for the correction of welding trajectories due to positioning errors using structured lighting and Gaussian T1 FLC.

The implementation results demonstrated the feasibility of the proposed Gaussian T1 FLC. The controller was validated experimentally with errors within the working range $\pm 20mm$ for the Y axis and $\pm 10mm$ in the Z axis. During experiments, the obtained results showed a maximum error of 1 mm, which is considered appropriate in practical manufacturing tasks.

The statistical analysis of the results indicated that there is statistical evidence to consider a mean of zero value for average error in both axes with $\alpha = 0.01$ for the triangular T1 FLC, while it shows that the Gaussian T1 FLC is more efficient since it works under a type system exploring the area more exhaustive using intermediate points. In addition, due to a greater number of diffuse rules and the use of Gaussian functions that include the value of the media and of the deviation standard, the model includes the uncertainties (noise) of the process sensors, i.e., the number of pixels detected during the processing of the image, which vary with the different levels of illumination.

Future work is looking into two areas: inspection of welding seam and type-2 fuzzy logic controllers. In terms of inspection is intended to expand the scope of the project by including the verification of width and height of the seam using a camera with greater dynamic range (approximately 120 dB), and the use of logarithmic algorithms that reduce high luminance that are present during the welding process.

Acknowledgment

Authors want to acknowledge the Consejo Nacional de Ciencia y Tecnología (CONACyT) for the scholarship provided to Ignacio Dávila-Ríos during his PhD studies, as well as the grant support from CONACyT No. 61373.

Bibliography

[1] Ignacio Davila-Rios, Ismael Lopez-Juarez, Gerardo M. Mendez, Roman Osorio-Comparan, Gaston Lefranc and Claudio Cubillos (2016); A Fuzzy Approach for on-line Error Compensation During Robotic Welding, *Computers Communications and Control (ICCCC), 2016 6th International Conference on*, ISBN 978-1-5090-1735-5, IEEE Xplore, DOI:10.1109/ICCCC.2016.7496766, 264- 270.

[2] Menno de Graaf , Ronald Aarts, Ben Jonker, Johan Meijer (2010); Real-time seam tracking for robotic laser welding using trajectory-based control, *Control Engineering Practice*, 18(8):944-953.

[3] Tero Santti, Jonne K. Poikonen, Olli Lahdenoja, Mika Laiho, Ari Paasio (2015); Online seam tracking for laser welding with a vision chip and FPGA enabled camera system, *IEEE International Symposium on Circuits and Systems (ISCAS), May 2015*, 1985 - 1988.

[4] Jinchao Liu, Zhun Fan, Sřren Ingvor Olsen, Kim Hardam Christensen, Jens Klćstrup Kristensen (2015); A Real-time Passive Vision System for Robotic Arc Welding, *IEEE International Conference on Automation Science and Engineering (CASE), Aug. 2015*, 389-394.

[5] T. Ngo, Y. Wang, T.L. Mai, M.H. Nguyen, J. Chen (2012); Robust Adaptive Neural-Fuzzy Network Tracking Control for Robot Manipulator, *International Journal of Computer Communications & Control*, ISSN 1841-9836, 7(2):341-352.

[6] I. Hartley and A. Zisserman (2004); *Multiple View Geometry in Computer Vision*, Cambridge University Press, ISBN: 0521540518, second edition, 2004.

[7] J. M. Mendel (2001); *Uncertain rule-based fuzzy logic systems: introduction and new directions*, Prentice-Hall, Upper Saddle River, NJ, 2001.

[8] John Yen, Reza Langari (1998); *Fuzzy Logic: Intelligence, Control and Information*, Prentice-Hall), ISBN 0135258170, 1998.

[9] Martín del Br io B, Sanz Molina A. (2006); *Redes Neuronales y Sistemas Borrosos*, 3rd edition, ISBN 978-84-7897-743-7, RA-MA, 2006.

[10] Montgomery D C. (2007); *Design and Analysis of Experiments*, John Wiley Sons, 6th edition, ISBN-10: 047148735X, 2007.

INTERNATIONAL JOURNAL OF COMPUTERS COMMUNICATIONS & CONTROL
ISSN 1841-9836, 12(2):217-226, April 2017.

Detecting Bridge Anaphora

D. Gîfu, L.I. Cioca

Daniela Gîfu*
1. Romanian Academy - Iaşi branch
Codrescu, 2 Iaşi, 700481, Romania
2. "Alexandru Ioan Cuza" University of Iaşi
General Berthelot St., 16, Iaşi, 700483, România
"Alexandru Ioan Cuza" University of Iaşi, România
*Corresponding author: daniela.gifu@info.uaic.ro

Lucian-Ionel Cioca
"Lucian Blaga" University of Sibiu
10, Victoriei Bd., Sibiu, 550024, România
lucian.cioca@ulbsibiu.ro

Abstract: The paper presents one of most important issues in natural language processing (NLP), namely the automated recognition of semantic relations (in this case, bridge anaphora). In this sense, we propose to recognize automatically, as accurately as possible, this type of relations in a literary corpus (the novel *Quo Vadis*), knowing that the diversity and complexity of relations between entities is impressive. Furthermore, we defined and classified the bridge anaphora type relations based on annotation conventions. In order to achieve the main goal, we developed a computational instrument, BAT (*Bridge Anaphora Tool*), currently still in a test (and implicitly an improvable) version. This study is intended to help especially specialists and researchers in the field of natural language processing, linguists, but not only.
Keywords: bridge anaphora, annotated novel, Bridge Anaphora Tool, testing corpus, corpus-driven, statistics.

1 Introduction

The novelty of this study consists in the development of a web application for the automated identification of bridge anaphora type relations in a corpus from the literary area. In this case, the target is the Romanian version of the novel *Quo Vadis*, authored by the Nobel laureate Henryk Sienkiewicz [24].

Initially, a similar study carried out by the same team consisted in the supervised extraction of Bridge Anaphora type relations, using WEKA statistics [8]. Moreover, there was defined a set of annotation conventions for 11 bridge relations as a result of manual annotations made by a team of trained students in Computational Linguistics.

The hypothesis of this paper is that the triggers have a fundamental role in the automated recognition of semantic relations generally and particularly of bridge anaphora relations.

The paper is structured in 5 sections. After a brief introduction about the importance of this study, section 2 mentions some important works focused on bridging anaphora. Section 3 describes bridge anaphora relations in the context of semantic relations and section 4 describes a new tool functionality, called BAT (*Bridge Anpahora Tool*). The last section highlights conclusions and mentions the future intentions, one of the main projects of Romanian researchers in NLP.

Copyright © 2006-2017 by CCC Publications

2 State of the art

In our context, the semantic relations play a fundamental role in the information extraction process [1], regardless of the nature of the corpus [2, 9, 10]. Up to now, researchers in the NLP area have allocated a lot of time to identifying the best annotation conventions of semantic relations for various literary types [3, 4, 12] based on which the process of automated recognition of semantic relations was not only simplified, but the accuracy of results increased as well, for example in their unsupervised extraction [7].

One of the best known studies on the "bridging" concept originates with H.C. Herbert [11]. He starts from several scenarios in which an inference step is needed in order to understand the sense intended by the speaker and he states that the text itself does not offer the solution for solving the inference relation; the reader (or the computational instrument/machine) must use his/its knowledge on the anaphora and the antecedent in order to make a correct text interpretation. In the automated recognition of semantic relations, a special attention is granted to the anaphora resolution [23, 25], using statistic models [19, 22], something that we too exploited.

NLP uses for recognizing entities and identifying relations in the text (bridging) systems based on manually created rules (see Hobb's algorithm) [15], but also systems using statistical models that are in turn based on automated learning techniques in order to lessen the workload, models such as Conditional Random Fields (CRF) [26].

3 Bridge anaphora in semantic relations context

In order to better understand a content, we need thinking instruments, necessary for discovering new ideas or for clarifying the existing ones, illustrating the link between them. The semantic relations [16] describe these interactions, that are indispensable for interpreting texts. The properties of semantic relations were described in [17], this marking the relations between two entities (called poles) as open class. The application describes 10 types of bridge anaphora[1].

3.1. A short introduction about semantic relations

The semantic relations are represented as being distributions over several paragraphs [18]. In processing the natural language, the semantic relations play a fundamental role in the field of Information Extraction (IE), that targets the automated extraction of structured information referring to entities such as person names, localities etc. from semi-structured or unstructured texts.

The ability to identify and understand these relations in a text can b useful in very many directions, such as: Machine Translation - MT; Computer Assisted Assessment - CAA; Clustering and so on.

In order to create an instrument that can carry out, for example the automated translation, the interpretation of anaphora is also very important, especially in cases in which the translation is from a language in which the pronouns have different forms for each gender, into a language in which the pronoun has the same form regardless of gender [15, 22].

[1] **Class-of** - relation between PERSON-CLASS & PERSON; **Has-as-member** - relation between PERSON-GROUP & PERSON; **Has-as-part** - relation between PERSON & PERSON-PART; **Has-as-subgroup** - relation between PERSON-GROUP & PERSON-GROUP; **Has-name** - relation between PERSON & PERSON-NAME; **ISA** - relation between PERSON & PERSON-CLASS; **Member-of** - relation between PERSON & PERSON-GROUP; **Name-of** - relation between PERSON-NAME & PERSON; **Part-of** - relation between PERSON-PART & PERSON; **Subgroup-of** - relation between PERSON-GROUP & PERSON-GROUP.

3.2. About bridge anaphora

Bridge anaphora [8] are referential semantic relations (beneath the co-referential or anaphoral ones) [5, 6] that includ linguistic expressions that give meaning to the analysed text (here, the narrative "thread" of the novel). Our documentation shows that the analysis of semantic relations is focused on structured corpuses such as: online newspapers, blogs, Wikipedia texts etc. [1].

A bridge anafora or "bridging" is a semantic relation that represents a link between the anaphora and the antecedent [11, 12]. These two elements will be mentioned in the following also as poles of a bridge-type semantic relation. In the next section we present the 10 types of bridge anaphora relations based on which the BAT was developed.

An example of bridge-type semantic relation:

Andrei este numit în diferite cercuri micuţul, din cauza înălţimii.

—>(En.) Andrei is called in different circles the little guy, because his height.
where:

- *Andrei* is an antecedent;
- *micuţul* —>(En.) *the little guy* is an anafor.

3.3. Bridge anaphora vs. anaphora

A bridge anaphora type relation differs from an anaphorical relation firstly by the fact that it can be identified in the text using a *trigger*. This trigger can be a word or a group of words that has the property of indicating the presence in the text of the bridge anaphora relation, helping to identify it.

In the following, we will exemplify the anaphorical relation and the bridge anaphora type relation in oder to clarify the difference between the two relations, both being referential type relations:

- Anaphorical relation (coreferential)

1:[Marcus] era foarte supărat pentru toate cele întâmplate în ultima perioadă, însă 2:[el] nu avea de gând să renunţe. —>(En.) 1:[Marcus] was very upset about what happened lately, but 2:[he] was not going to give up.
=>[2] anaphorical relation [1];

- Bridge anaphora type relation (below, the type **class-of**[2])

Cândva, 1:[Petronius] fusese guvernator în 2:[Bitinia]... —>(En.) Sometime 1:[Petronius] was governor in 2:[Bithynia]...
=>[1] bridge anaphora type relation [2], while governor in is the trigger for this relation.

This is a segmentation annotation in XML standoff format:

```
<W LEMMA="cândva" MSD="Rg" POS="ADVERB" id="1" offset="0">Cândva</W>
<W LEMMA="," MSD="COMMA" id="2" offset="6">,</W>
<CLAUSE CONTINUE="27" ID="CLAUSE31">
<ENTITY ID="E000900036" TYPE="PERSON">
```

[2]**Class-of** - is a bridge anaphora type relation linking a PERSON-CLASS type concept to a PERSON type instance.

```
<REFERENTIAL FROM="E000900036" ID="REF000900582" TO="E000700030" TYPE="coref">
<W Case="oblique" Definiteness="no" EXTRA="NotInDict" Gender="feminine"
LEMMA="Petronius" MSD="Npfpon" Number="plural" POS="NOUN" Type="proper" id="3"
offset="8">Petronius</W>
</REFERENTIAL>
</ENTITY>
</CLAUSE>
<W EXTRA="intranzitiv" LEMMA="fi" MSD="Vmil3s" Mood="indicative"
Number="singular" POS="VERB" Person="third" Tense="long" Type="predicative"
id="4" offset="18">fusese</W>
<ENTITY ID="E000900037" TYPE="PERSON">
<W Case="direct" Definiteness="no" Gender="masculine" LEMMA="guvernator"
MSD="Ncmsrn" Number="singular" POS="NOUN" Type="common" id="5"
offset="25">guvernator</W>
</ENTITY>
<W LEMMA="în" MSD="Sp" POS="ADPOSITION" id="6" offset="36">în</W>
<CLAUSE CONTINUE="31" ID="CLAUSE32">
<ENTITY ID="E000900038" TYPE="LOCATION">
<REFERENTIAL FROM="E000900038" ID="REF000900584" TO="E000900036"
TYPE="class-of">
<REFERENTIAL FROM="E000900038" ID="REF000900584" TO="E000800035" TYPE="coref">
<W EXTRA="NotInDict" LEMMA="Bitinia" MSD="Np" POS="NOUN" Type="proper" id="7"
offset="39">Bitinia</W>
</REFERENTIAL>
</REFERENTIAL>
</ENTITY>
</CLAUSE>
```

The anaphorical relations are a widely debated subject [12, 13, 14], proven by numerous specialty papers that present computational instruments for the automated identification of these relations, especially for the pronominal anaphora [15, 22]. This type of relation is much easier to identify in the text, as opposed to a bridge anaphora type semantic relation, because both poles of the the relation, the anaphora and the antecedent refer to the same entity [20]. In order to be able to automatically identify bridge type anaphorical relations, there is necessary a more complex mechanism, that would carry aut in a first phase a preprocessing of the text for its de-ambiguization that consists in segmentation, tokenization, lemmatization, part-of-speech tagging, name entity recognition, and anaphora resolution.

4 BAT - description

Bridge Anaphora Tool is a computational instrument implemented in Java language, on the framework Java Server Faces and uses a series of libraries[3]. BAT is created for the automated recognition of bridge type semantic relations, more precisely of the 10 types of referential relations for which annotation conventions have been determined.

The output XML file was used in the process of training and testing. We chose the novel Quo Vadis [24], given that it is a corpus translated into more than 40 de languages, having an impressive number of entities and semantic relations. Using the instrument PALinkA [21]

[3]see http://primefaces.org/

the entities and semantic relations were annotated manually. The annotator was already used successfully for annotating the novel Quo Vadis, a work presented in [3].

This web application (fig. 1) executes in a first phase the training process after which the automated recognition can be initiated.

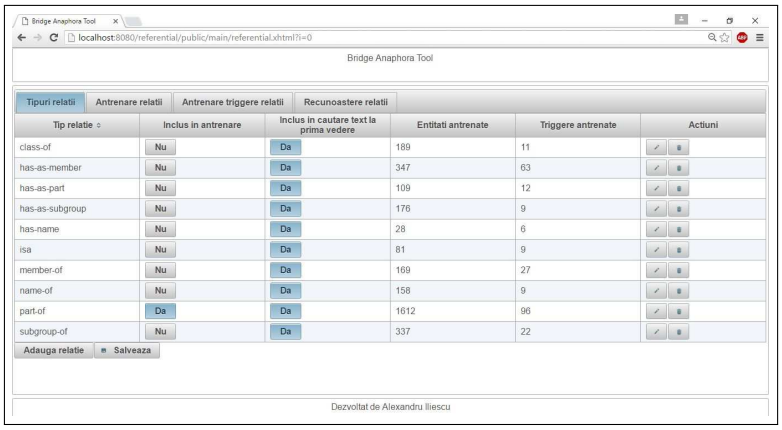


Figure 1: The interface of the computational tool

In the following, we describe briefly the work methodology. For the trening process, following steps were conceived:

- The option "YES" is selected for the relations that will be included in the training;

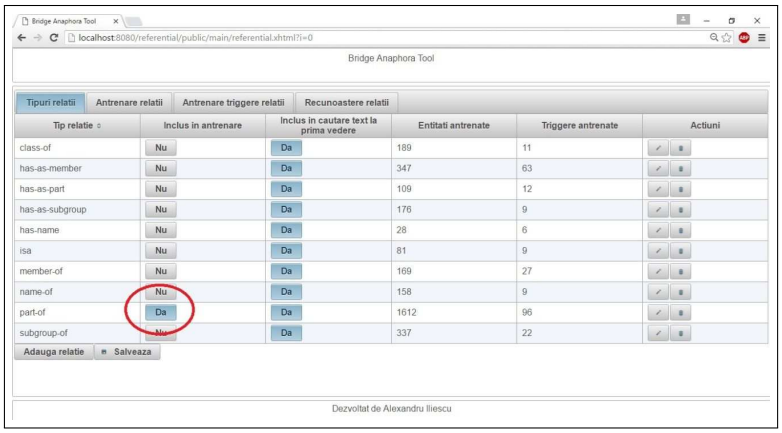


Figure 2: BAT - selecting relations for training

- The XML file is loaded from the application, using the button "Train relations", the XML is selected (the manually annotated corpus) after which the button "Process file" is pressed.

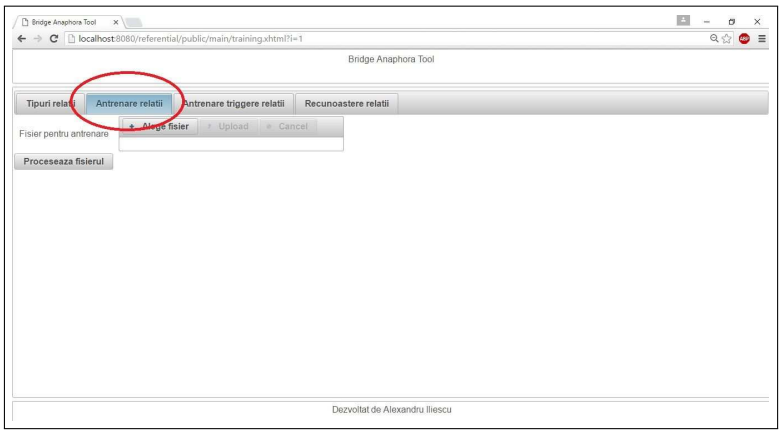


Figure 3: BAT - menu for introducing the corpus and for trening relations

When the BAT identifies the tag <REFERENTIAL>, it carries out four steps:
- it saves in the MySQL database the type of relation (it can be one of the 10 "class-of", "has-as-member", etc.) in the table "referential_type";

Figure 4: BAT - the table "referential_type" generated after the training

- it saves the type of entity from the identified relation (in the example above we have TYPE="PERSON") in the table "entity_type";

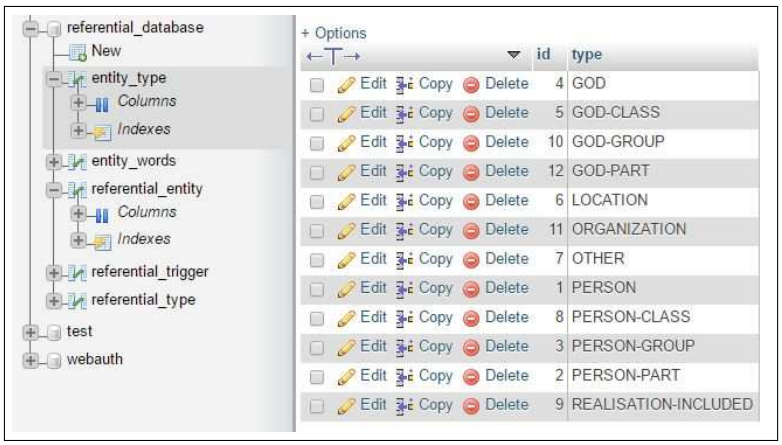


Figure 5: BAT - the table "entity_type" generated after the training

- it saves the word/words from the tag <ENTITY>specific to the relation in the table "entity_words"; if there are two or more words, they are concatenated with the symbol "|";

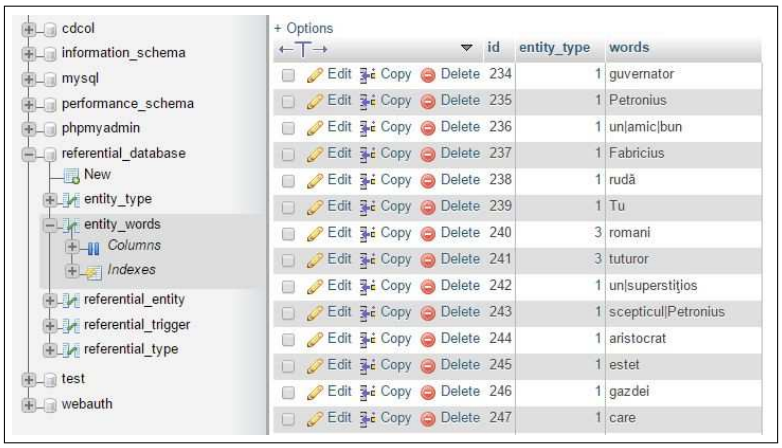


Figure 6: BAT - the table "entity_words" generated after training

- it saves the actual structure of a bridge type semantic relation, i.e. the "TYPE" and the words forming it, in XML they being identifiable with the elements "ID", "FROM" and "TO".

Figure 7: BAT - the table "referential_entity" generated after training

The processing of the XML file in the training phase of the BAT for one or more bridge type semantic relations can take from one minute to several hours, function of the number of relations existing in the annotated corpus. For the "part-of" relation, the training took 2.67 hours, being the most often encountered in the XML file, with a number of 1612 relations.

5 Statistics and interpretations

Bridge Anaphora Tool used for training 66% of the corpus of Quo Vadis.

Mitkov (1998) suggested, for measuring the performance of a computational instrument aiming at identifying anaphorical relations in the text, an equation that defines its success rate.

The definition of the success rate is as follows:

BAT success rate = 534 correctly identified relations / 1035 total existing relations = 61.5%.

So at this moment, the BAT recognized correctly over 61% of the bridge anaphora type semantic relations that should have been identifying in the text, thus fulfilling the set goal.

Table 1: The results of recognizing the bridge type semantic relations with BAT

Bridge Anaphora types	Bridge Anaphora number identified with BAT in driven corpus	Bridge Anaphora number identified automatically in testing corpus	Bridge Anaphora number identified manually in testing corpus
class-of	189	58	28
has-as-member	347	115	82
has-as-part	109	31	12
has-as-subgroup	176	55	22
has-name	28	9	7
isa	81	25	14
member-of	169	51	38
name-of	158	53	29
part-of	1612	530	249
subgroup-of	337	108	53
Total	**3206**	**1035**	**534**

We think that the variations of the values in the column "number of relations identified automatically by BAT in the testing corpus" are due also to the fact that the instrument searches "mechanically" in the preprocessed text rigid definitions of the relations.

For example: entity of the type PERSON-NAME + PERSON =>relation "name-of".

Moreover, there exist two relations that have the same definition, namely the relations: *has-as-subgroup* and *subgroup-of* being given by the entities of type PERSON-GROUP+PERSON-GROUP, the only difference between them being made by the triggers, during testing.

Conclusions and future work

This paper presents a methodology for the automated recognizing of 10 bridge anaphora (or bridging) type semantic relations, each having several particularities. The achieved results are promising, offering a base for future researches. We suggest using in parallel of machine learning models (Naïve Bayes and Support Vector Machines).

The BAT instrument, developed for the automated recognition of Bridge Anaphora relations, will be improved through the addition of several triggers to the existing list, or in the situation in which there would be available even more data for training.

BAT is far from being a perfect instrument, but it can be improved since it showed to be efficient at least for an experimental purpose for various applications in the NLP area.

Acknowledgments

We thank Alexandru Iliescu for developing the BAT instrument. We are also grateful to all colleagues from NLP-Group@UAIC-FII who developed the tools for natural language processing used in this research.

Bibliography

[1] Banko, M., Cafarella, M. J., Soderland, S., Broadhead, M., Etzioni, O. (2007); Open information extraction from the Web, *IJCAI07 Proceedings of the 20th international joint conference on Artifical intelligence*, 2670-2676.

[2] Buraga, S.C., Cioca, M., Cioca, A. (2007); Grid-based decision support system used in disaster management, *Studies in Informatics and Control*, 16(3):283-296.

[3] Cristea, D., Gîfu, D., Diac, P., Maraunduc, C., Bibiri, A., Scutelnicu, A., Colhon, M. (2014); *Quo Vadis: A Corpus of Entities and Relations*, Springer International Publishing Switzerland, 2014.

[4] Cristea, D., Dima, G. E., Postolache, O. D., Mitkov, R. (2002); Handling complex anaphora resolution cases, *Proceedings of the Discourse Anaphora and Anaphor Resolution Colloquium, Lisbon, 2002*, 1-6.

[5] Branco, A., McEnery, T., Mitkov, R. (2005); *Anaphora Processing: Linguistic, Cognitive and Computational Modelling*, John Benjamins, 2005.

[6] Cojocaru, L. (2000); The study of the anaphoric relationship on the Romanian corpora, Inference in Computational Semantics, at the Inference in Computational Semantics, *ICoS-2, Schloss Dagstuhl, Germany, July 29-30.*

[7] Conrath, J, Afantenos, S., Asher, N., Muller, P. (2014); Unsupervised extraction of semantic relations using discourse cues, *International Conference on Computational Linguistics - COLING (Dublin, Ireland)*, 2184-2194.

[8] Gîfu, D., Iliescu, A. (2014); Analysis of Bridge Anaphora across novel, *Procedia- Social and Behavioral Sciences*, 180: 1474-1480.

[9] Gîfu, D.; Cioca, M. (2013); Online civic identity. Extraction of features, *Procedia - Social and Behavioral Sciences*, 76:366-371.

[10] Gîfu, D.; Cioca, M. (2014) Detecting Emotions in Comments on Forums, *International Journal of Computers Communications & Control*, 9(6):694-702.

[11] Herbert, H. C. (1975); Bridging. *Proceedings of the 1975 Workshop on. Theoretical Issues in Natural Language. Processing, TINLAP '75*, 169-174.

[12] Hendrickx, I., Clercq, O., Hoste, V. (2011); Analysis and reference resolution of bridge anaphora across different text genres, *Anaphora Processing and Applications - 8th Discourse Anaphora and Anaphor Resolution Colloquium, DAARC 2011, Faro, Portugal*, LNAI 7099:1-11.

[13] Krahmer, E., Piwek, P. (1997); Varieties of anaphora, *Proceedings of the 11th Amsterdam Colloquium, University of Amsterdam*, 5-20.

[14] Korzen, I., Buch-Kromann, M. (2006); Anaphoric relations ĂŽn the Copenhagen Dependency Treebanks, *Proceedings of COLING-ACL 06*, 3:83-98.

[15] Lappin, S., Leass, H. J. (1994); An Algorithm for Pronominal Anaphora Resolution, *Computational Linguistics*, 20(4):535-561.

[16] Liang, T., Wu, D. (2010); Automatic Pronominal Anaphora Resolution ĂŽn English Texts, *Computational Linguistics and Chinese Language Processing*, 9(1):21-40.

[17] Murphy, M. L. (2003); *Semantic relations and the lexicon Antonymy, synonymy, and other paradigms*, Cambridge University Press, Cambridge, UK, 2003.

[18] Năstase, V., Nakov, P., Seaghdha, D. O., Szpakowicz, S. (2013); *Semantic relations between nominals*, California: Morgan & Claypool Publishers, 2013.

[19] Niyu, G., Hale, J., Charniak, E. (1998); A Statistical Approach to Anaphora Resolution, *Proceedings of the Sixth Workshop on Very Large Corpora, Montreal, Canada*, 161-170.

[20] Nedoluzhko, A., Mirovsky, J., Ocelak, R., Pergler, J. (2009); Extended Coreferential Relations and Bridging Anaphora in the Prague Dependency Treebank, *Proceedings of the 7th Discourse Anaphora and Anaphor Resolution Colloquium (DAARC 2009), Goa, India*, 1-16.

[21] Orăşanu, C. (2003); PALinkA: A highly customisable tool for discourse annotation, *Proceedings of the 4 th SIGdial Workshop on Discourse and Dialogue, ACL'03*, 1-5.

[22] Rello, L., Ilisei, I. (2009); A comparative study of spanish zero pronoun distribution, *Proceedings of the International Symposium on Data and Sense Mining, Machine Translation and Controlled Languages (ISMTCL)*, 1-5.

[23] Wasow, T. (1967); *Anaphoric relations in English*, Ph.D. dissertation, Massachusetts Institute of Technology, 1967.

[24] Sienkiewicz, H. (1991); *Quo Vadis*, translated in Romanian by Luca, R. and Linţă, E., Ed. Tezi, Bucharest.

[25] Singh, S., Lakhmani, P., Mathur, P., Morwal, S. (2014); Analysis of Anaphora Resolution System for English Language, *International Journal on Information Theory*, 3(2):5157, DOI : 10.5121/ijit.2014.3205.

[26] Žitnik, S, Šubelj, L, Bajec, M (2014); SkipCor: Skip-Mention Coreference Resolution Using Linear-Chain Conditional Random Fields, *PLoS ONE*, 9(6): e100101. doi:10.1371/journal.pone.0100101.

INTERNATIONAL JOURNAL OF COMPUTERS COMMUNICATIONS & CONTROL
ISSN 1841-9836, 12(2):227-237, April 2017.

The Model for Learning Objects Design Based on Semantic Technologies

D. Gudoniene, R. Maskeliunas, D. Rutkauskiene

Daina Gudoniene*, Rytis Maskeliunas,
Danguole Rutkauskiene
Kaunas University of Technology
Studentu str. 50, Kaunas, 51392, Lithuania
*Corresponding author: daina.gudoniene@ktu.lt
rytis.maskeliunas@ktu.lt, danguole.rutkauskiene@ktu.lt

Abstract: The paper presents a comparison of state of the art methods and techniques on implementation of learning objects (LO) in the field of information and communication technologies (ICT) using semantic web services for e-learning. The web can serve as a perfect technological environment for individualized learning which is often based on interactive learning objects. This allows learners to be uniquely identified, content to be specifically personalized, and, as a result, a learner's progress can be monitored, supported, and assessed. While a range of technological solutions for the development of integrated e-learning environments already exists, the most appropriate solutions require further improvement on implementation of novel learning objects, unification of standardization and integration of learning environments based on semantic web services (SWS) that are still in the early stages of development. This paper introduces a proprietary architectural model for distributed e-learning environments based on semantic web services (SWS), enabling the implementation of a successive learning process by developing innovative learning objects based on modern learning methods. A successful technical implementation of our approach in the environment of Kaunas University of Technology is further detailed and evaluated.
Keywords: learning object, semantic technologies, web applications, learning environments.

1 Introduction

A variety of tools and systems could be used for the implementation of learning activities through e-learning processes. When talking about e-learning technologies the most important and painful question we get is: what method should be used for the integration of interactive learning objects (LO) in the development of e-learning? The painful thing about this question is that a usual questioner has often been misled to believe that there is only a single tool that does everything that everybody needs to be done: to create, host, and access e-learning material. Successful e-learning strategies and scenarios may require integration of dozens of software products chosen from hundreds of candidates sprawling across multiple categories: development of learning objects, delivery of knowledge, content management, communication and collaboration, live learning and assessment, etc. They can also be categorized according to the possibilities of implementation of curriculum (realization of learning events: imitate, receive information, exercise, explore, experiment, create, self-reflect, debate); technological properties (e.g. synchronous, asynchronous, web based, PC application, mobile application, open source, free service); application domain (language learning, intercultural competences, ICT skills, time management skills, study habits skills, etc.). However, independent on the purpose or functionality, all tools and systems are often integrated in the virtual learning environment (VLE), which can reflect the discipline by providing a well-designed, visually stimulating interface that genuinely supports the needs of real world learning. A web-based education method suits this by providing more

Copyright © 2006-2017 by CCC Publications

flexibility and intelligence. However, recent developments in the area of Semantic Web, while contributing to the solution to these problems, also raise new issues that must be solved [3].

The aim of the research in this paper is to present an architectural model of distributed e-learning environments based on semantic web services (SWS) with a goal to improve the effectiveness of a learning process by introducing innovative learning objects that combine variety of multimedia elements and other learning material as well as structural integrity of advanced learning methods. Extraction of semantic relations has always been a challenging problem in multimedia data. Various architectures of multimedia databases have been developed in the past but the need to refine them still remains in order to get the desired results of users' interest, to extract semantics from multimedia data in a way the user perceives them.

2 Related works

Interactive learning possibilities were improved with the introduction of internet technologies; however, it still fails to reach the full potential. With new, more interactive internet technologies there is even more to be captured and adopted, such as the public knowledge contained in blogs, wikis, social bookmarking services, social networks, etc. [8, 17]. The ontologies, the Semantic Web, and the Social Semantic Web offer a new perspective on intelligent educational systems by providing intelligent access to and management of Web information, and semantically richer modelling of applications and their users [2-3]. This allows supporting more accurate representations of learners, their learning goals, learning materials and contexts of use, as well as more efficient access and navigation through learning resources with a primary aim to advance intelligent educational systems to achieve improved e-learning efficiency, flexibility, and adaptation for both single users and communities of users.

The notion of the Social Semantic Web describes an emerging design approach for building Semantic Web applications that employs Social Software techniques. Social Semantic Web systems typically elicit domain knowledge through semi-formal ontologies, taxonomies or folksonomies.

Techniques related to educational content could be provided in different forms [1, 9, 18] and environments that have relations with semantic web services. As web-based education, it has become a very important branch of educational technology [2, 7, 19]. Classroom independence and platform independence of web-based education, availability of authoring tools for developing web-based courseware, cheap and efficient storage and distribution of course materials, hyperlinks to suggested readings, digital libraries, and other sources of references relevant for the course are but a few of a number of clear advantages of Web-based education. By analysing design of learning objects and essential characteristics of a range of proven learning activities, we can generate a set of requirements for the IS architecture. For example, a proven existing learning activity based on implementation of learning objects might enable students to work simultaneously across a network on a design tool, such as a graphics program, and share the results in separate windows. This learning activity therefore generates the computational requirement to allow the use of any shareable application this way.

Other researchers [6, 9, 10] have analysed a number of digital resources to be used and reused for learning (learning objects) and concluded that a number of those is constantly increasing. Therefore, description of learning objects with metadata is important as it allows enhancing search, retrieval and usage of learning objects and because it is very important in the integration process into any environment to efficiently organize a training process. Learning objects can be considered not only as resources providing affiliated materials, but also as methodological resources which include teachers' experiences, reflections, examples or instructions of usage of content objects and descriptions of learning methods [9, 18]. However, existing standards and specifications for learning objects metadata are not intended for including methodological

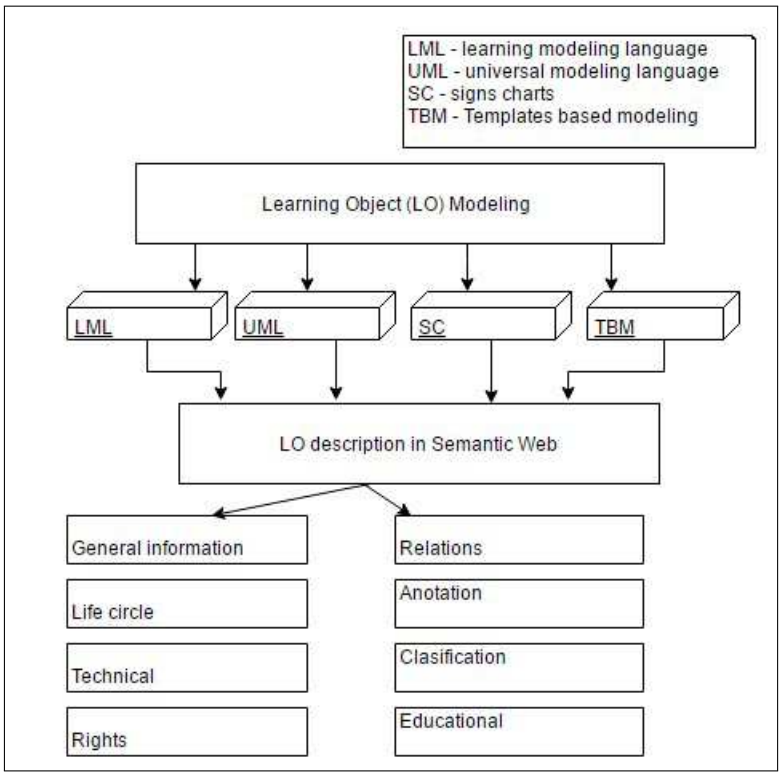


Figure 1: Principal schema of learning objects relations

resources and learning method descriptions together with content objects.

Modeling of learning objects and search of learning objects of the same content in the semantic web is a challenge to every researcher. The metadata and ontologies (see Fig. 1) let to find and to reuse learning objects in different situations as they make LO machine-understandable. According to Rehman and Kifor [21], ontologies are like repositories: they are helpful in exchange of knowledge, reusing existing knowledge, for reasoning and inferring on existing knowledge.

Dagiene et al. are discussing that digital learning resources by themselves are not as valuable as their target application in the learning process as well as a properly selected learning method. Therefore, metadata is indispensable and semantic web services should be intended part of the design of learning objects.

The systematically prepared corpus of metadata helps all the parties involved in learning process to cooperate, use and share learning resources [16], [13].

Interface of a learning object usually presents the same content and has the same look and feel for every user, regardless of a student's learning needs and individual characteristics. Some interfaces are "customizable"; the user can choose to modify some characteristics of the graphical user interface. However, this does not entirely satisfy the needs of educational content presentation to the users. Aspects of content personalization, proper interaction and efficient presentation are also important.

Some research on a design of learning objects has been already conducted [4]-[6] and some researchers [18] made the effort to combine learning objects and constructivism focus largely on how learning objects can be used in specific constructivist learning environments instead of building a universal generic structure.

The results of research presented by Gurbuz et al. [5] suggest that a new architectural model for learning objects development must be designed with a common login system that provides

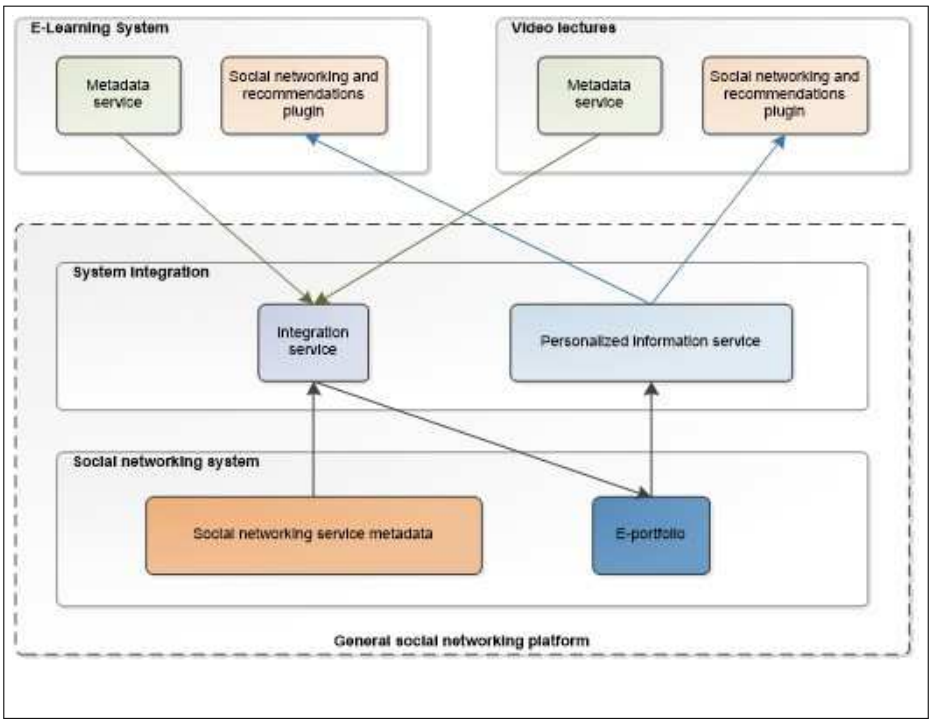


Figure 2: Architectural model of the platform

information about a person trying to connect (for example, name, address, email address, user code).

Aroyo at al. [10] and Gote et al. [7] have found a key to enable the interoperability and to capitalize on the semantic conceptualization and ontologies, common standardized communication syntax, and large-scale service-based integration of educational content and functionality provision, and usage based on a model of Semantic Web Services for Education.

The central role in achieving unified authoring support plays the process-awareness of authoring tools which should reflect the semantic evolution of e-learning systems.

Gurbuz et al. [5] present a system architecture model which is designed and based on service-oriented architecture technology (SOA). This architecture intended to provide an ability to create a flexible system focused on the provision of services with the ability to implement effective services and systems solutions for an integration of standard learning objects into information systems.

The platform (see architectural model of the platform in Fig. 2) can be structured in the following logical levels [10]-[12]:

- User level, consisting of user and system interfaces, implemented according to user specification and system requirements;
- The level of External systems; the adapter is provided for managing Web services. In addition, a system of components is required for external inclusion;
- The data level involves all data management, sanctioning, monitoring, archiving and storage components.

When connected to one of the web services by the use of other systems the second time logging is not necessary anymore as the system automatically checks the permission of online users and

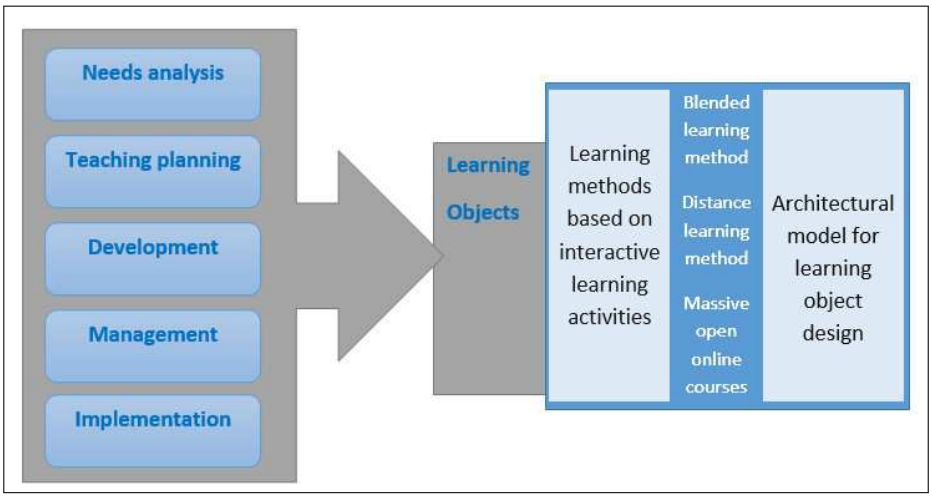


Figure 3: Integration of learning objects into interactive learning courses

authorizes them. The architectural model provides main functionality necessary for the design and implementation of learning objects.

A central role of achieving unified authoring support plays the process-awareness role of authoring tools which should reflect the semantic evolution of e-learning systems [14], [15], [20]. The researchers are working on the novel architecture of learning objects to solve the problem of unification of learning objects and to start working on standardization of LOs and to offer templates to users to design content based learning objects for e-learning systems related to semantic web services technologies.

The semantic relations with the learning objects are well constructed and to be provided from learning objects' interface.

3 A proprietary model of building LOs for semantic e-learning environments (SWS)

We propose a novel architecture for designing modern learning objects that can be easily applied to various domains.

A model is based on identified key learning objects phases that make an influence on the organization of study process as well as on finding technological solutions to design LOs and to develop technical implementation. We have carried out a comprehensive data analysis on the implementation of e-learning processes starting with the requirement specification to the delivery phase. In the evaluation, we have analysed the design of learning objects, methods of learning (i.e. blended learning method, distance learning method, etc.) and methods of providing massive open online courses.

The architectural model (see Fig. 4) was designed theoretically (based on the state of the art and best practice scenarios); its implementation started on a step-by-step basis. One of the requirements was to include metadata of each section that could be used to provide some useful information about the task goal, keywords, background, etc. Going up to the knowledge granularity level (see Fig. 3), multiple sections can form a component. Multiple components related to a learning topic can be grouped. When a learning object is generated by an instructor/author, the expert's experience can be embedded into the configuration of the course, such as selection of section contents, selection of components, sequence of components, display modes of components, etc. When it is generated by the learning environment, certain organizational

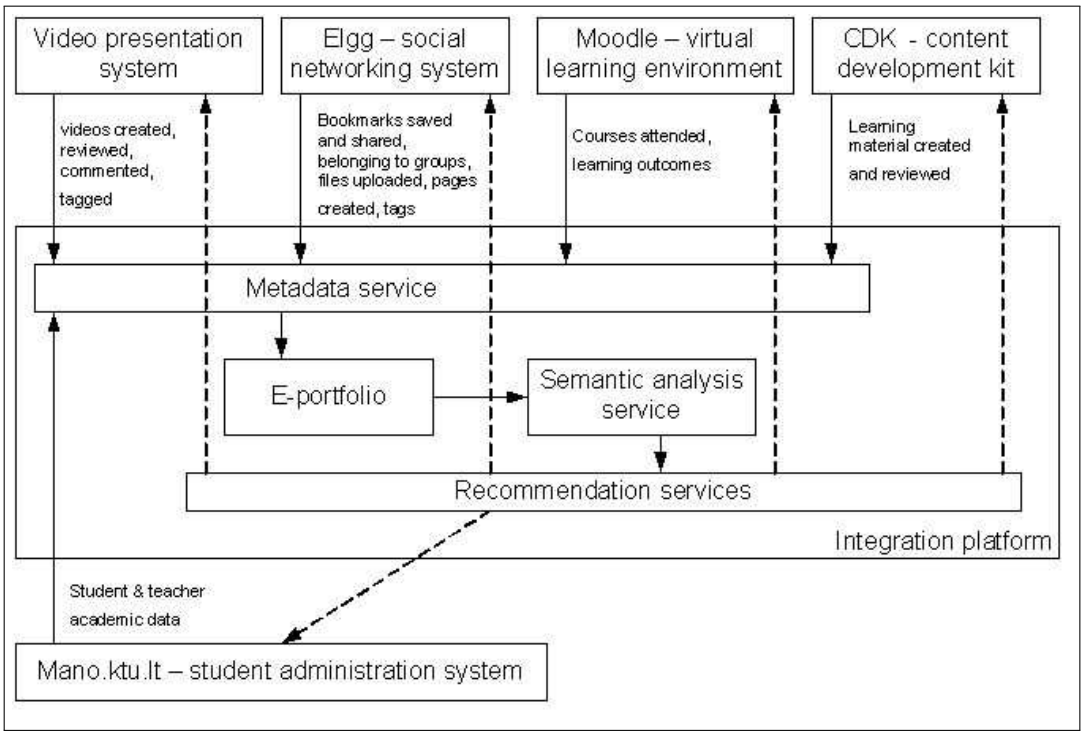


Figure 4: Architectural model for the design and delivery of learning objects

patterns can be applied to generate a course corresponding to the learner's profile and to the learning goal.

The architecture is based on SOA principles. The functions will be provided as services; the internal service processes are intended to optimize the use of the provision of direct procedure calls that can be realized directly in the server application procedures. Some kind of activities can be performed as services and will be directly integrated into learning objects.

Through the components of Web Services (WS), there is a possibility to take an advantage of the VLE and its functionality which provides a specific service such as data recording, getting a report, etc. Moreover, the service component of our system can be used to work in a workflow. The workflow can be used for both systems and services of VLE IS provided by external integrated systems.

IS and all its components are adapted using standard IT platforms, operating [13], [16] systems and existing computer network infrastructure. Information system can provide non-formal education and implementation of learning objects as well as self-tests carried out in the content management system that assures this essential functionality and the relations by semantic web services directly to content of learning objects, including preparation of learning objects, metadata description and semantic relations to other open resources; development of self-tests, integration into the software facilities; learning objects, learning programs and courses in preparation of and public access to its management; IS and user administration.

High resolution digital objects are stored in our university's VLE IS repository at open.ktu.lt. These objects must then be converted (transformed) into different formats and quality facilities for further implementation and content creation of different types (for different training programmes and courses).

IS and it's components are designed to work in the environment which was adapted to communicate with standard IT platforms, operating systems and the computer network.

Looking to the IS on student interface we can view learning objects in different ways. One

way is to click "view default" to choose to view the course created by the author/instructor if such a version is available. Alternatively, students can choose and have the course generated automatically by the system according to some patterns like some external teaching methodology. There is a way to have a choice to view the learning objects sorted by the profile by which the learning materials will be generated session by session. Finally, students can choose to view all of the raw components related to the topic.

Our approach demonstrates the possibility of using the constructivism learning theory to guide the design of learning objects based upon the original prototype. The collaboration among authors of LO's is supported, while learners can also actively participate in the construction of learning objects. It provides a way to allow learners to grasp the whole picture of the course quickly. The ease of viewing contents of learning objects iteratively in different ways assists learners to learn efficiently in constructivist learning environments.

Through the Web Services (WS) components of the system, users and designer can take an advantage of the VLE and its functionality which provides specific services, such as initiating business processes to record data to get a report and so on. Besides, the service component of the system can be used to work in the workflow. Workflow can be used for both VLE IS systems and services provided by external integrated systems.

Some parts of the described system are under constant improvement. Changes in system functionality processes have to be implemented and adapted to the system. Our proposed process modelling makes it easier to adapt to the emerging new needs, for instance, to decide to adjust the sequence of necessary steps to abandon any of the steps to change the conditions of implementation steps.

The workflow can be easily incorporated during the steps carried out by the external system via services (service), as well as allow to easily changing the step executor. For example, a workflow step is carried out by an external system which presents the results of the VLE from that system.

IS functions will assure the insertion of educational content (including different format of learning objects), testing, and content development: development of digital material and metadata description; design of learning objects, design of courses and open access management; development of tests and assignments and relation to the targeted program; virtual learning environment and users administration.

Key property of the Semantic Web architecture (common-shared-meaning, machine-processable metadata), enabled by a set of suitable agents, seems to be powerful enough to satisfy the e-learning requirements: fast, just-in time and relevant learning. Learning material is semantically annotated and it may be easily combined in a new learning course in case of a new learning demand [23]. The process is based on semantic querying and navigation through learning materials enabled by the ontological background. In fact, the Semantic Web could be treated as a very suitable platform for implementation of e-learning objects since it provides all means for (e-learning) ontology development, ontology-based annotation of learning materials, their composition in learning courses and (pro)active delivery of the learning materials through e-learning portals.

In our system, learning objects are distributed on the web but they are also linked to agreed network of ontologies. This enables construction of user-specific courses by allowing semantic querying for various topics of interest. Software agents of the Semantic Web can be used to enable co-ordination between other system agents and proactive delivery of learning materials in the context of actual problems.

The Semantic Web can be as decentralised as possible. This enables an effective co-operative content management. Content is determined by an individual user's needs and aims to satisfy the needs of every user. The users by using personalised agent searches for learning material will

be customised for her/his needs. The ontology is the link between user needs and characteristics of the learning material.

The Semantic Web enables the use of knowledge, provided in various forms, regarding semantic annotation of content.

Distributed nature of the Semantic Web enables continuous improvement of learning materials. It enables the use of distributed knowledge provided in various forms, enabled by semantic annotation of content. Distributed nature of the Semantic Web enables continuous improvement of LO.

4 Practical realization of the designed architectural model for learning objects development

The developed architectural model for the design of learning objects includes content development tools and social network integration and video lecturing system. The content acquisition and control measures were implemented by using the open-source virtual learning environment Moodle, an open source social networking system ELGG and open-source Drupal CMS integrated with a system for video lecture recording as well as the tool CKD dedicated to the development of learning objects. The system allows users to create / edit / delete content, search, use e-learning content and share with other users for evaluation. Facility installed and configured user roles and rights connected allow to form a single login system within active web services, in order to produce the information and allow for service centres to carry out the instructions sent by the inclusive media.

Semantic tools have been implemented to assist the workflow of course creation delivery, to revise a recommendation process of a relevant content and people in the context of the course and the institution, to assist students by recommending resources that match the topics of their assignments and personnel that may be able to support their activities, to help in group formation for collaborative work based on students' background, personal preferences and successful prior collaboration, to add support for critical thinking and argumentation by visualizing arguments and linking relevant discussions.

Implementation of semantic technologies allowed enabling integration, searching and matching of nodes of information. Large university repositories, like triple stores, where information can be efficiently stored, searched and managed, were also improved by a more efficient semantic model.

According to the architectural model presented in the section 2, changes were made to IS distance learning management system "mano.ktu.lt".

The platform was created with the idea to aggregate the information distributed by different institutional departments including personal user's needs (portfolio). This system data of user activity is gathered by metadata service from different e-learning systems. Metadata is then structured and saved as a user e-portfolio. Semantic analysis service is activated on user e-learning event, for example, the user is logging to any of the platform's systems. Semantic analysis service gets e-portfolio data and performs semantic analysis using several methods: user-based Collaborative Filtering, cosine similarity, person correlation, jaccard - tanimoto index, Sorensen coefficient.

Recommendations for learning materials and other users (for promotion of collaboration) are presented to the active user after an analysis is performed in the plugin of the environment where the user is performing his learning activities. For example, if a user is in a video presentation system, s/he gets similar videos based on the analysis of all data on all systems together (semantic proximity between users in elgg, moodle, cdk, mano.ktu.lt environments is analysed) – if two

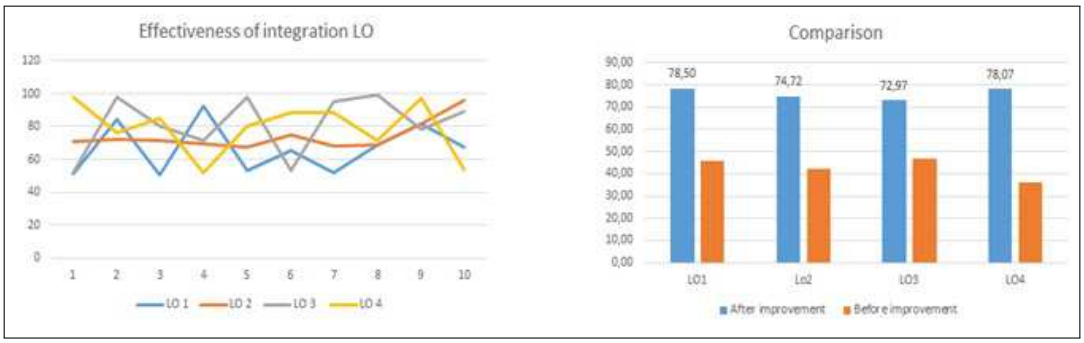


Figure 5: Effectiveness of LO integration

and more users collaborate in the same elgg group, they might want to check the same videos on a video presentation system.

5 Initial experimental evaluation

As there is no way to effectively introduce objective measures for other than performance tasks, a specified survey was issued to conceive the effectiveness of current milestone version. The users were requested to answer the questionnaire and express their opinion on the design and delivery of learning objects as well as about practical implementation of LOs into the institutional system: if it is friendly, easy to use or not.

The research was implemented on national institutional level where 94 authors of academic courses were requested to answer the questionnaire. 54 % of respondents were aged between 45-60 and 46 % were aged between 30-45. The respondents were requested to answer the main question which technological solutions dedicated to design learning objects were friendlier to use? From an age perspective, we can claim that age was not a very important factor and the respondents were interested in learning this novel architectural model and were coping easily with a standard template to design interactive learning object that does not require special competences to provide an interactive learning content to students.

We have also asked to compare the creation of standard learning objects vs non-standard learning objects used in educational process. The result showed that the users were friendlier with the standard objects (see a comparison of usability of learning objects based on the architectural model and not in Table 1) where the number of respondents provided data on the LO usability in the architectural platform.

The experimental research on the effectiveness of LO integration (see Fig.5) was carried out by means analysing data on the platform developed on the architectural model (see Fig. 4) and using of high resolution learning objects to be provided for the study process. In other cases the supplied video objects have proven as a more effective way to work with students who were in different locations, and to leave video records for self-learning as well. This is true even for the development and delivery of massive open online courses in our university.

According to the respondents, successful integration of architectural model into practice opened an easy way to teachers and lecturers to make video records of lectures.

After tuning a model on the remarks of our experimental participants, a component of video conferencing system was modified into a more user-friendly experience (VIPS, http://vips.liedm.lt) was developed. IS environment became very friendly to use due to a novel combination of modern technologies and established methods of pedagogy.

Conclusions

New approaches of the learning objects modelling plays the central role in achieving unified authoring support in the process-awareness of authoring tools which should reflect the semantic evolution of e-learning systems. However, the researchers are working on the architecture on LO design to solve the problem of the unification of learning objects and to start standardization of LOs and to suggest the templates for users designing content based standard learning objects for e-learning systems based on semantic web services technologies.

For conclusions the authors identify that the developed model offers perfect technology and environment for individualized learning based on interactive learning objects not only for teachers but for learners as well, as they can be uniquely identified, content can be specifically personalized, and the learning progress can be monitored, supported and assessed.

The ontologies, the Semantic Web, and the Social Semantic Web offer a new perspective on intelligent educational systems by providing intelligent access to and management of Web information, and semantically richer modelling of applications and their users.

The Semantic Web enables the use of distributed knowledge provided in various forms, provided by semantically annotated content. Distributed nature of the Semantic Web enables continuous improvement of learning objects.

The research data shows that architecture for the design of learning objects can be applied to various domains and that authors can easily work with in order to design and integrate of learning objects. At the same time, the technological solution for architectural model based on semantic web services cannot be influenced by the age of users as this is not very important and the respondents were interested in architectural model and standard template to design interactive learning object having very easy technologies that do not require special competences to provide to students interactive learning content.

Bibliography

[1] Ermalai, I.; Dragulescu, B.; Ternauciuc, A.; Vasiu, R. (2013); Building a Module for Inserting Microformats into Moodle, *Advances in Electrical and Computer Engineering*, 13(3): 23-26.

[2] Kruk, S.; Gzella, A.; Dobrzanski, J.; McDaniel, B.; Woroniecki, T. (2007); E-Learning on the Social Semantic Information Sources, *Lecture Notes in Computer Science*, 4753: 172-186.

[3] Ermalai, I.; Mocofan, M.; Onita, M.;Vasiu, R. (2009); Adding Semantics to Online Learning Environments, *Proceedings of 5th International Symposium on Applied Computational Intelligence and Informatics – SACI2009*, 569-573.

[4] Targamadze,A.; Balbieris, G.; Kubiliunas, R. (2005); The new generation of virtual learning environments in Lithuania, *Information Technology and Control*, 34(3): 276-284.

[5] Gurbuz, T.; Gudoniene, D.; Rutkauskiene, D. (2013); System architecture model based on service-oriented architecture technology, *Proceedings of 19th international conference Information and Software Technologies*, 102-113.

[6] Kurilovas, E.; Dagiene, V. (2009); Multiple Criteria Comparative Evaluation of E-Learning Systems and Components, *Informatica*, 20(4): 499-518.

[7] Grtze, P.; Engelund, C.; Mortensen, R.L.; Paszkowski, S. (2009); Cross-National Interoperability and Enterprise Architecture, *Informatica*, 20(3): 369-396.

[8] Abdul Hamid, O.; Abdul Qadir, M.; Iftikhar, N.; Ur Rehman, M.; Uddin Ahmed, M.; Ihsan, I. (2007); Generic Multimedia Database Architecture Based upon Semantic Libraries, *Informatica*, 18(4): 483-510.

[9] Dagiene, V.; Jevsikova, T.; Kubilinskiene, S. (2013); An Integration of Methodological Resources into Learning Object Metadata Repository, *Informatica*, 24(1): 13-34.

[10] Aroyo, L.; Dicheva, D. (2013); The New Challenges for E-learning, *The Educational Semantic Web Educational Technology & Society*, 7 (4): 59-69.

[11] Rutkauskiene, D.; Gudoniene, D.; Cibulskis, G.; Suk, O. (2013); ICT architecture for online learning approach, *Proceedings of Information and software technologies - ICIST 2012*, 373-387.

[12] Rutkauskiene, D.; Mark, R.; Kubiliunas, R.; Gudoniene, D. (2013); Functional architecture of a service-oriented integrated learning environment, *Proceedings ECEL 2013 - 12th European Conference on e-learning*, 431-439.

[13] Kuzucuoglu, A. E.; Gokhan, E. (2011); Development of A Web-Based Control And Robotic Applications Laboratory For Control Engineering Education,*Information Technology and Control*, 40(4): 352-358.

[14] Bajec, M. (2008); A Framework and Tool-Support for Reengineering Software Development Methods, *Informatica*, 19(3): 321-344.

[15] Bersin, J. (2008); Social Networking and Corporate Learning, *Certification Magazine*, 10: 14-14.

[16] Fertalj, K.; Hoic-Bozic, N.; Jerković, H. (2010); The Integration of Learning Object Repositories and Learning Management Systems, *Computer Science and Information Systems*, 7: 387-407.

[17] Grodecka, K.; Wild, F.; Kieslinger, B. (2011); *How to Use Social Software in Higher Education*, Poland.

[18] Targamadze, A.; Petrauskiene, R. (2010); Impact of Information Technologies on Modern Learning, *Information Technology and Control*, 39(3): 169-175.

[19] Vitiutinas, R.; Silingas, D.; Telksnys, L. (2011); Model-driven plug-in development for UML based modelling systems, *Information technology and control*, 40(3): 191-201.

[20] Alsultanny, Y. (2010); e-Learning System Overview based on Semantic Web, *The Electronic Journal of e-Learning*, 4 (2): 111-118.

[21] Rehman, Z.; Kifor, S. (2014); A Conceptual Architecture of Ontology Based KM System for Failure Mode and Effects Analysis, *International Journal of Computers Communications & Control*, 9 (4): 463-470.

INTERNATIONAL JOURNAL OF COMPUTERS COMMUNICATIONS & CONTROL
ISSN 1841-9836, 12(2):238-253, April 2017.

WEBIRA - Comparative Analysis of Weight Balancing Method

A. Krylovas, N. Kosareva, E.K. Zavadskas

Aleksandras Krylovas, Natalja Kosareva*,
Edmundas Kazimieras Zavadskas
Vilnius Gediminas Technical University
Sauletekio al. 11, LT-10223 Vilnius, Lithuania
aleksandras.krylovas@vgtu.lt, natalja.kosareva@vgtu.lt,
edmundas.zavadskas@vgtu.lt
*Corresponding author: natalja.kosareva@vgtu.lt

Abstract: The attributes weight establishing problem is one of the most important MCDM tasks. This study summarizes weight determining approach which is called WEBIRA (WEight Balancing Indicator Ranks Accordance). This method requires to solve complicated optimization problem and its application is possible by carrying out non trivial calculations. The efficiency of WEBIRA and other MCDM methods – SAW (Simple Additive Weighting) and EMDCW (Entropy Method for Determining the Criterion Weight) compared for 4 different data normalization methods. The results of the study revealed that more sophisticated WEBIRA method is significantly efficient for all considered numbers of alternatives. Efficiency of all methods decreases with increasing number of alternatives, but WEBIRA is still applicable, while application of other methods is impossible as the number of alternatives is greater than 11. WEBIRA is the least affected by the data normalization, while EMDCW is the most affected method.
Keywords: WEBIRA, SAW, EMDCW, multi-attribute decision making (MADM), entropy, KEMIRA.

1 Introduction

From the large diversity of MCDM methods some are very simple to use methods, and the other – complex, requiring more effort and computing resources. This article analyses the attributes weighting task, which is solved by different methods. One of the main and well known multiple criteria decision making (MCDM) methods is calculation of weighted averages of the the performance values of alternatives evaluated in terms of attributes (criteria):

$$S^{(j)}(W,R)=\sum_{i=1}^{n} w_i r_i^{(j)},\ j=1,2,\ldots,m, \tag{1}$$

here $W=(w_1,w_2,\ldots,w_n)$, $0\leqslant w_i\leqslant 1$, $\sum\limits_{i=1}^{n} w_i=1$ is vector of weights, n – number of attributes, $R=\begin{pmatrix} r_1^{(1)} & r_2^{(1)} & \ldots & r_n^{(1)} \\ r_1^{(2)} & r_2^{(2)} & \ldots & r_n^{(2)} \\ \ldots & \ldots & \ldots & \ldots \\ r_1^{(m)} & r_2^{(m)} & \ldots & r_n^{(m)} \end{pmatrix}$ – matrix of alternatives $j\in\{1,2,\ldots,m\}$ estimates, which elements $0\leqslant r_i^{(j)}\leqslant 1$ are obtained by certain measurements or expert assessments applying different calculation procedures, which are usually referred to as normalization methods (see [1, 2]). For example, in this article maximum method (Max), sums method (Sum), minmax method (MinMax) and vector normalization (Vctr) will be used.

The mentioned MCDM method is known as WSM (Weighted Sum Model). It was noticed by Churchman in 1954 [3] that "course of action that maximizes the expected total weighted

Copyright © 2006-2017 by CCC Publications

efficiency (effectiveness) is optimum". This method proposed for solving MCDA problems in [4]. Zionts and Wallenius in 1976 [5] emphasysed that the basis of this method is the intuitive understanding that the "overall utility function is assumed to be implicitly a linear function, and more generally a concave function of the objective functions". Triantaphyllou in 2000 [6] noticed that in the maximization case, the best alternative is the one that yields the maximum total performance value (1). Though the method is rather old it is still relevant and frequently used, the articles with its applications appearing in solid academic journals nowadays. The article [7] proposes a modified version of the Weighted Sum Model that takes into account decision-maker preferences and provides a possibility of higher interactivity in the selection of the most suitable alternative. A mathematical model for a Dynamic Weighted Sum Method (DWSM) is presented in [8]. In Hwang and Yoon [9] this method was called Simple Additive Weighting (SAW). The name stuck, and today one can find a number of articles in which this method is actually called as SAW. Triantaphyllou in [10] drew attention to aggregating of benefit and cost criteria in four different MCDA methods. In [11] Simple Additive Weighting is proposed as a metamodel for other MCDA methods. Generalized SAW under fuzzy environment and the relative preference relation proposed in [12] to easily and quickly solve FMCDM problems. A new comprehensive overview of multiple attribute decision-making techniques and their applications is presented in [13]. In the review [14] of MCDM literature it was noted that the Analytic Hierarchy Process (AHP) in the individual tools and hybrid MCDM in the integrated methods were ranked as the first and second methods in use.

MacCrimmon in [15] noticed that 1) as the number of relevant attributes and alternatives increases, the ability of the decisionmaker to handle the problem decreases; 2) using a combination of MCDM methods frequently may be more feasible than using any one method separately. A study [16] presents a hybrid MCDM method combining SAW, Techniques for Order Preference by Similarity to an Ideal Solution (TOPSIS) and Grey Relational Analysis (GRA) techniques. The ranking results show that multiple MCDM methods are more trustworthy than those generated by a single MCDM method. A new COmbinative Distance-based ASsessment (CODAS) method to handle MCDM problems is proposed in [17]. To improve the accuracy of weighted sum and weighted product models (WSM and WPM), in [18] the Weighted Aggregated Sum Product Assessment (WASPAS) method was applied as an aggregation operator on WSM and WPM. In the paper [19], an extended version WASPAS-IVIF method is proposed which can be applied in uncertain decision making environment. In [20] authors revealed that MCDM methods work fairly well in estimating the number of clusters in a data set.

Another aspect of the MCDM methods – weight coefficients W selection, which can be accomplished by using *a priori* information (expert's estimates, subject specific knowledge) or *a posteriori* information of matrix R itself. The latter are sometimes called objective weight determining methods [21].

In the articles [22–25] a new weight determining approach which is applicable to the tasks when matrix R is composed of two or more components is proposed. All MCDM methods using this methodology can be assigned to the group of weights balancing methods, which hereinafter we call WEBIRA (WEight Balancing Indicator Ranks Accordance).

The main idea of this approach is maximizing compatibility of the two (or more) sets of attributes which are treated as independent. Optimization task is being solved and criteria weights are sought throughout the weight balancing procedure. WEBIRA so far have not been compared with other MCDM methods, thus relevant is the question when it is appropriate to use. WEBIRA is suitable for a very important economic benefit carrying tasks, it has both scientific and practical meaning. For example, developers constructing sustainable products and technologies must pay attention to three main components such as economic development, social development and environmental protection. MCDM problems of sustainability could be solved by applying

WEBIRA for 3 sets of attributes – economic, social and environmental components of sustainable development.

A major criticism of MADM is that different techniques may yield different results when applied to the same problem. It does not exist multiple criteria evaluation method which is "best under any circumstances". Therefore the comparative analysis of various MCDM methods determining which method is best for a particular case is relevant and important task. The performance of eight methods: ELimination and Choice Expressing REality (ELECTRE), TOPSIS, Multiplicative Exponential Weighting (MEW), SAW, and four versions of AHP investigated in simulation experiment in [26]. Simulation parameters are the number of alternatives, criteria and their distribution. In general, all AHP versions behave similarly and closer to SAW than the other methods. ELECTRE is the least similar to SAW. The following performance order of methods was established: SAW and MEW (best), followed by TOPSIS, AHPs and ELECTRE. The comparative analysis of MCDA methods SAW and COPRAS (Complex Proportional Assessment) describing their common and diverse characteristics is proposed in [27]. The paper [28] presents an empirical application and comparison of six different MCDM approaches (between them WSM, AHP, TOPSIS, COPRAS) for the purpose of assessing sustainable housing affordability. TOPSIS, SAW, and Mixed (Rank Average) for decision-making as well as AHP and Entropy for obtaining the weights of attributes have been compared in [29]. Mixed method as compared to TOPSIS and SAW is the preferred technique, moreover, AHP is more acceptable than Entropy for weighting. The comprehensive study [30] carried out the comparative analysis among well-known and widely-used methods WPM, WSM, TOPSIS, AHP, PROMETHEE, ELECTRE, when applied to the reference problem of the selection of wind turbine support structures for a given deployment location. The outcomes of this research highlight that more sophisticated methods, such as TOPSIS and Preference Ranking Organization METHod for Enrichment Evaluation (PROMETHEE), better predict the optimum design alternative.

WEBIRA method requires to solve complicated optimization task and therefore it relates to executing non trivial computer calculations. Naturally, the question arises as to when it would appear reasonable to apply WEBIRA, and when – the less sophisticated approaches. In this article Monte Carlo-type experiments are performed and WEBIRA is compared with two simple objective weight determining methods: AVRG – the simple arithmetic average, i. e. the weighted sum with equal weights and the Entropy Method for Determining the Criterion Weight (EMDCW) described in [21]. The Shannon entropy method [31] is one of the most famous approach for determining the objective attribute weights. Entropy measures the uncertainty associated with a random variable, i.e. the expected value of the information transmitted to the decision maker. The authors of the paper [21] have combined the best features of the entropy method and the CILOS (the Criterion Impact Loss) approach to obtain a new method – Integrated Determination of Objective CRIteria Weights, or (IDOCRIW). In [32] three methods have been used for estimating criteria weights: Entropy, CILOS and IDOCRIW, while for the selection of priority well-known and widely used MCDM methods SAW, TOPSIS and COPRAS have been used in MCDM analysis of operating of rotor systems with tilting pad bearings.

Another problem the article dealt with – the comparative analysis of efficiency of some data normalization methods. A state-of-the-art survey on the influence of normalization techniques in ranking is proposed in [33]. Thirty-one methods were identified, classified and evaluated for use in materials selection problems. Review of normalization methods used in construction engineering and management, and their applications presented in [34].

The article is organized as follows. In Chapter 2 the algorithm of solving the optimization problem and the case study of its application is proposed. In Chapter 3 random matrices generating scheme is described. In Chapter 4 the transformation formulas for matrix of estimates proposed. Chapter 5 describes the process of numerical experiments and methods of efficiency

comparison. Chapter 6 describes the statistical analysis of the results of numerical experiments, summarizes and proposes recommendations for application of various MCDM methods and normalizing procedures.

2 Algorithm of WEBIRA method

Suppose, that matrix R is composed of two components $R = (P|Q)$, $P = \left(p_i^{(j)}\right)_{m\times n_p}$, $Q = \left(q_i^{(j)}\right)_{m\times n_q}$, $n_p + n_q = n$.

Two weighted sums are calculated

$$S_P^{(j)} = \sum_{i=1}^{n_p} w_{Pi} p_i^{(j)},\ S_Q^{(j)} = \sum_{i=1}^{n_q} w_{Qi} q_i^{(j)},\ j = 1, 2, \ldots, m. \tag{2}$$

Coefficients $W_P = \left(w_{p1}, w_{p2}, \ldots, w_{pn_p}\right)$, $W_Q = \left(w_{q1}, w_{q2}, \ldots, w_{qn_q}\right)$ satisfy monotonicity conditions:

$$1 \geqslant w_{p1} \geqslant w_{p2} \geqslant \cdots \geqslant w_{pn_p} \geqslant 0, \quad 1 \geqslant w_{q1} \geqslant w_{q2} \geqslant \cdots \geqslant w_{qn_q} \geqslant 0. \tag{3}$$

Inequalities (3) are set from expert estimates, when k experts line up attributes p_{ie}, q_{ie} according to their importance:

$$p_{i_1 e} \succ p_{i_2 e} \succ \cdots \succ p_{i_{n_p} e},\ q_{i_1 e} \succ \cdots \succ q_{i_2 e} \succ q_{i_{n_q} e},\ e = 1, 2, \ldots, k. \tag{4}$$

When we have *a priori* information about experts evaluations (4), inequalities (3) could be obtained by different methods. In the article [22] Kemeny median has been adapted for this purpose and the name KEMIRA (KEmeny Median Indicator Ranks Accordance) proposed for the method. In the articles [23–25] inequalities (3) were set by calculating entropy values or by application of voting theory methods. As the inequalities (3) indicating the weight preferences are established, all of the mentioned methods can be assigned to the objective weight determining, because there is not need of further information to set the weights of the attributes. We consider the task when inequalities (3) already established and the weights W_P, W_Q determining task is being solved. The task is formulated as minimization of a certain distance or measure:

$$s\left(W_P, W_Q\right) = \min_{W_P, W_Q} \sqrt{\frac{1}{m}\sum_{j=1}^{m}\left(S_P^{(j)} - S_Q^{(j)}\right)^2}, \tag{5}$$

where the weights W_P, W_Q are satisfying the inequalities (3). So, all the mentioned MCDM methods [22–25] can be assigned to the group of weight balancing methods, which we call WEBIRA.

In this article we analyze only the benefit type attributes, i.e. whose higher value is better. When optimization task (5),(3) is already solved, the ranks can be assigned to the alternatives $j \in \{1, 2, \ldots, m\}$ depending on the size of the weighted sums $S_P^{(j)}$ and $S_Q^{(j)}$. Suppose that j_1^P, j_2^P, …, j_m^P, j_1^Q, j_2^Q, …, j_m^Q are such numbers of alternatives that the weighted sums (2) are satisfying the inequalities:

$$S_P^{j_1^P} \geqslant S_P^{j_2^P} \geqslant \cdots \geqslant S_P^{j_m^P}, \quad S_Q^{j_1^Q} \geqslant S_Q^{j_2^Q} \geqslant \cdots \geqslant S_Q^{j_m^Q}. \tag{6}$$

Denote a set of the best k alternatives according to the first n_p attributes $A_k^P = \{j_1^P, j_2^P, \ldots, j_k^P\}$, according to the last n_q attributes – $A_k^Q = \{j_1^Q, j_2^Q, \ldots, j_k^Q\}$ and their intersection $A = A_k^P \cap A_k^Q$.

The meaning of the sets A_k^P and A_k^Q is selecting the best k alternatives according to the attributes P and Q respectively while the meaning of the set A – the best alternatives according to the both attributes. The purpose of WEBIRA method is to balance weights so that a number of elements $|A|$ of the set A would be not less than the certain number. So, it is required to find a sufficient number of the best alternatives according to both attributes P and Q. In the articles [22–25] the minimizing tasks (5),(3) have been solved together with $\max|A|$ under various parameter k values of the sets A_k^P and A_k^Q. In the current article we limit ourselves to the case $k = 1$ and require $A \neq \emptyset$, i. e. we will search the only one the best alternative according to the both attributes P and Q.

This additional condition can be formulated as follows:

$$S_P^{(j_0^P)} = \max_{j=1,2,\ldots,m}\{S_P^{(j)}\}, \quad S_Q^{(j_0^Q)} = \max_{j=1,2,\ldots,m}\{S_Q^{(j)}\},\ j_0^P = j_0^Q. \tag{7}$$

Algorithm of solving the problem (5),(3),(7) is as follows.

1. By random re-selection among weights W_P, W_Q satisfying conditions (3) the weights satisfying an additional condition (7) are being searched.

2. If the weights W_P^0, W_Q^0 were not found after $iter_0$ iterations, algorithm is finishing work and it is concluded that the weights can not be determined.

3. If the weights W_P^0, W_Q^0 are set, the loss value (5) is fixed as s^0, directions ΔW_P, ΔW_Q are selected at random and the new weights $W_P^1 = W_P^0 + h\Delta W_P$, $W_Q^1 = W_Q^0 + h\Delta W_Q$ calculated, here h is the predetermined value.

4. The correction of weights W_P^1, W_Q^1 is performed as follows $W \to \tilde{W}$:
   $\tilde{W} = (\tilde{w}_1, \tilde{w}_2, \ldots, \tilde{w}_n)$, $\tilde{w}_i = \begin{cases} 1, \text{if } w_i \geqslant 1, \\ 0, \text{if } w_i \leqslant 0, \\ w_i, \text{else}, \end{cases}$
   then the weights are normalized $w_i = \frac{\tilde{w}_i}{\sum\limits_{i=1}^{n} \tilde{w}_i}$.

5. Checking whether the new weights W_P^1, W_Q^1 satisfy the condition (7). If they satisfy and the number of iterations does not exceed the established limit $iter_1$ the algorithm moves to the Step 7.

6. If the number of iterations exceeds the established limit $iter_1$, algorithm finishes its work with the determined weights W_P^0, W_Q^0.

7. The loss value (5) is calculated $s^1\left(W_P^1, W_Q^1\right)$. If $s^1 \leqslant s^0$, we substitute $W_P^0 = W_P^1$, $W_Q^0 = W_Q^1$, $s^0 = s^1$ and then the new weights W_P^1, W_Q^1 are calculated with the same directions ΔW_P, ΔW_Q. Go to the weight correction procedure to the Step 4.

8. If $s^1 \geqslant s^0$, the directions ΔW_P, ΔW_Q are changed randomly, i. e. go to the Step 3 of algorithm.

Note. Algorithm parameters $iter_0$, $iter_1$ and h are set empirically and their values were set respectively to 225, 200 and 0.05.

Example. Provide the case study of the described algorithm application.
Set parameter values: $m = 6$, $n_x = 5$, $n_y = 4$, $iter_0 = 225$, $iter_1 = 100$, $h = 0.05$,

$$P = \begin{pmatrix} 0.8333 & 1.0000 & 0.6667 & 0.8000 & 0.3333 \\ 0.6667 & 0.2000 & 0.6667 & 1.0000 & 1.0000 \\ 0.8333 & 0.8000 & 1.0000 & 0.2000 & 0.1667 \\ 1.0000 & 0.8000 & 0.8333 & 0.6000 & 1.0000 \\ 0.1667 & 0.4000 & 0.3333 & 1.0000 & 0.3333 \\ 0.3333 & 0.2000 & 1.0000 & 0.6000 & 0.8333 \end{pmatrix},$$

$$Q = \begin{pmatrix} 1.0000 & 1.0000 & 0.3333 & 0.4000 \\ 0.3333 & 0.6000 & 0.6667 & 0.8000 \\ 0.3333 & 0.4000 & 0.3333 & 0.6000 \\ 0.3333 & 1.0000 & 0.3333 & 1.0000 \\ 0.5000 & 0.6000 & 0.5000 & 0.8000 \\ 0.1667 & 0.6000 & 1.0000 & 1.0000 \end{pmatrix}. \tag{8}$$

The initial point: $(N = 1)$ $W_P = (0.5556, 0.1111, 0.1111, 0.1111, 0.1111)$,
$W_Q = (0.3636, 0.3636, 0.1818, 0.0910)$, $S_P = (0.7741, 0.6889, 0.7037, 0.9148, 0.3222, 0.4778)$,
$S_Q = (0.8242, 0.5333, 0.3818, 0.6363, 0.5636, 0.5515)$. The ranks of the alternatives: $Rangs_P = (4, 1, 3, 2, 6, 5)$, $Rangs_Q = (1, 4, 5, 6, 2, 3)$. We see, that condition (7) is not satisfied, i. e. two criteria P and Q differently determine the best alternative.

Let's skip some of the checked weights W_P, W_Q and provide only some of the calculations results: $(N = 119)$ $W_P = (0.3478, 0.3478, 0.1739, 0.0870, 0.0435)$,
$W_Q = (0.2500, 0.2500, 0.2500, 0.2500)$. $S_P = (0.8377, 0.5478, 0.7667, 0.8667, 0.3565, 0.4478)$,
$S_Q = (0.6833, 0.6000, 0.4166, 0.6666, 0.6000, 0.6917)$. The ranks of the alternatives are: $Rangs_P = (4, 1, 3, 2, 6, 5)$, $Rangs_Q = (6, 1, 4, 5, 2, 3)$ and again the condition (7) is not satisfied.

The initial point was found when $(N = 169)$, $W_P = (0.5000, 0.5000, 0.0000, 0.0000, 0.0000)$, $W_Q = (0.3636, 0.3636, 0.1818, 0.0910)$, $S_P = (0.9166, 0.4334, 0.8167, 0.9000, 0.2833, 0.2667)$, $S_Q = (0.8242, 0.5333, 0.3818, 0.6363, 0.5636, 0.5515)$.

The ranks of the alternatives: $Rangs_P = (1, 4, 3, 2, 5, 6)$, $Rangs_Q = (1, 4, 5, 6, 2, 3)$ already responding the condition (7).

The initial loss calculated by the formula (5) is $s^0 = 0.660953$. Randomly determined directions: $\Delta_P = (0.30, -0.30, -0.22, -0.13, -0.05)$, $\Delta_Q = (-0.41, 0.31, -0.20, -0.08)$. The function (5) is increasing in this direction, therefore, the opposite direction was chosen and the weights were set as follows: $W_P^1 = W_P^0 - h \cdot \Delta_P$, $W_Q^1 = W_Q^0 - h \cdot \Delta_Q$. Then the weights were adjusted (Step 4 of algorithm). The minimum value $s^1 = 0.653260$ of the function (5) in the direction Δ_P, Δ_Q was obtained with the weights $W_P = (0.4900, 0.4900, 0.0111, 0.0067, 0.0022)$, $W_Q = (0.3772, 0.3416, 0.1882, 0.0930)$.

The second iteration of the algorithm (random change in direction) is as follows. The direction vectors are: $\Delta_P = (0.35, 0.24, -0.22, -0.16, 0.04)$, $\Delta_Q = (-0.40, 0.39, -0.20, 0.00)$. In this case the function declines in the direction $W_P^1 = W_P^0 - h \cdot \Delta_P$, $W_Q^1 = W_Q^0 - h \cdot \Delta_Q$ and the minimum value $s^1 = 0.648413$ of the function (5) is obtained when the weights are $W_P = (0.4813, 0.4813, 0.0224, 0.0147, 0.0003)$, $W_Q = (0.3932, 0.3188, 0.1962, 0.0918)$.

We present some further iterations results:
$Iter = 17$, $s^1 = 0.575767$, $W_P = (0.4384, 0.4321, 0.0779, 0.0482, 0.0033)$,
$W_Q = (0.3942, 0.3456, 0.1468, 0.1134)$.

$Iter = 29$, $s^1 = 0.508748$, $W_Q = (0.4119, 0.4057, 0.0832, 0.0804, 0.0188)$,
$W_Q = (0.4094, 0.4035, 0.0935, 0.0935)$.
$Iter = 65$, $s^1 = 0.450759$, $W_P = (0.3447, 0.3447, 0.1488, 0.1488, 0.0130)$,
$W_Q = (0.4454, 0.4380, 0.0583, 0.0583)$.
$Iter = 93$, $s^1 = 0.388068$, $W_P = (0.2453, 0.2453, 0.2453, 0.2453, 0.0187)$,
$W_Q = (0.3954, 0.3954, 0.1227, 0.0866)$.

Notice that $Iter_1 = 100$ iterations were accomplished, but we failed to reduce the value $s^1 = 0.388068$ of the loss function, i. e. in all randomly selected directions the function (5) increased.

Given the weights obtained in the 93-th iteration the values of criteria are
$S_P = (0.8158, 0.6402, 0.6982,\ 0.8119, 0.4723, 0.5389)$,
$S_Q = (0.8662, 0.5201, 0.3828, 0.6546, 0.5655, 0.5124)$.

The ranks of the alternatives: $Rangs_P = (1, 4, 3, 2, 6, 5)$, $Rangs_Q = (1, 4, 5, 2, 6, 3)$.

Please note that for the algorithm realization it was necessary to apply a relatively complicated computer program which was realized in C++. However, one can easily check the calculations and this may be done in each step of the algorithm independently of other steps.

3 Random matrices estimates generation

The elements $x_i^{(j)}$, $y_i^{(j)}$ of the estimates matrices $X = \left(x_i^{(j)}\right)_{m\times n_x}$ and $Y = \left(y_i^{(j)}\right)_{m\times n_y}$ are the integers simulating the scores of the expert estimates $x_i^{(j)} \in \{1, 2, \ldots, b_i^X\}$, $y_i^{(j)} \in \{1, 2, \ldots, b_i^Y\}$ of the alternatives $i \in \{1, 2, \ldots, m\}$. Each row of matrices X and Y corresponds to one alternative

$$X = \begin{pmatrix} x_1^{(1)} & x_2^{(1)} & \cdots & x_{n_x}^{(1)} \\ x_1^{(2)} & x_2^{(2)} & \cdots & x_{n_x}^{(2)} \\ \cdots & \cdots & \cdots & \cdots \\ x_1^{(m)} & x_2^{(m)} & \cdots & x_{n_x}^{(m)} \end{pmatrix}, Y = \begin{pmatrix} y_1^{(1)} & y_2^{(1)} & \cdots & y_{n_y}^{(1)} \\ y_1^{(2)} & y_2^{(2)} & \cdots & y_{n_y}^{(2)} \\ \cdots & \cdots & \cdots & \cdots \\ y_1^{(m)} & y_2^{(m)} & \cdots & y_{n_y}^{(m)} \end{pmatrix}. \tag{9}$$

The columns of matrices (9) arranged in descending order of attributes priorities. The first line of matrix is generated with preset probabilities P_{ik}^X, P_{ik}^Y:

$$\begin{aligned} &P\{x_i^{(1)} = b_i^X - k\} = P_{ik}^X,\ k = 0, 1, 2, \ldots, b_i^X - 1, \\ &P\{y_i^{(1)} = b_i^Y - k\} = P_{ik}^Y,\ k = 0, 1, 2, \ldots, b_i^Y - 1. \end{aligned} \tag{10}$$

Antecedent probabilities P_{ik}^X, P_{ik}^Y chosen in such way, that the first alternative should have on average higher estimates. Other alternatives estimates generated with the equal probabilities:

$$\begin{aligned} &P\{x_i^{(j)} = b_i^X - k\} = \tfrac{1}{b_i^X},\ k = 0, 1, \ldots, b_i^X - 1,\ i = 1, 2, \ldots, n_x, \\ &P\{y_i^{(j)} = b_i^Y - k\} = \tfrac{1}{b_i^Y},\ k = 0, 1, \ldots, b_i^Y - 1,\ i = 1, 2, \ldots, n_y, \\ &j = 2, 3, \ldots, m. \end{aligned} \tag{11}$$

Therefore, the second and all other alternatives are treated as a kind of noise making heavy recognition of the first – the best alternative. The more alternatives we have, the more difficult is the task of identification.

The experiments were carried out with the following parameter values:
$b_i^X = 6$, $i = 1, 2, 3, 4, 5$, $b_i^Y = 6$, $i = 1, 2, 3, 4$. Probabilities of the first alternative estimates:

$$P\{x_{1,2}^{(1)} = 6\} = P\{x_{1,2}^{(1)} = 5\} = 0.5;\ P\{x_{1,2}^{(1)} = l\} = 0., l = 1, 2, 3, 4.$$

$$P\{x^{(1)}_{3,4,5}=6\}=0.;\ P\{x^{(1)}_{3,4,5}=l\}=0.25, l=2,3,4,5;\ P\{x^{(1)}_{3,4,5}=1\}=0.$$

Probabilities of other alternatives estimates are equal:

$$P\{x^{(j)}_i=l\}=\frac{1}{6},\ j=2,3,\ldots,m,\ i=1,2,3,4,5,\ l=1,2,3,4,5,6.$$

Similarly selected estimates probabilities of the first and other alternatives according to Y:

$$P\{y^{(1)}_{1,2}=6\}=P\{y^{(1)}_{1,2}=5\}=0.5;\ P\{y^{(1)}_{1,2}=l\}=0., l=1,2,3,4.$$

$$P\{y^{(1)}_{3,4}=6\}=0.;\ P\{y^{(1)}_{3,4}=l\}=0.25, l=2,3,4,5;\ P\{y^{(1)}_{3,4}=1\}=0.$$

$$P\{y^{(j)}_i=l\}=\frac{1}{6},\ j=2,3,\ldots,m,\ i=1,2,3,4,\ l=1,2,3,4,5,6.$$

The example of generated matrix with the best first alternative:

$$(X|Y)=\left(\begin{array}{ccccc|cccc} 5&6&3&4&2&6&5&4&4\\ 4&4&4&2&3&1&2&5&3\\ 3&4&5&1&3&2&2&5&5\\ 3&3&6&6&5&6&4&4&4\\ 4&3&4&3&5&6&5&6&4\\ 2&1&5&5&2&4&6&3&1\\ 1&2&3&2&2&2&5&6&4\\ 4&2&2&3&1&6&2&5&6 \end{array}\right). \tag{12}$$

4 Transformations of estimates matrix

Recall that randomly generated matrices X, Y elements $x^{(j)}_i$, $y^{(j)}_i$ are the integers while input data of WEBIRA method – matrices P, Q elements acquire real values in the interval $[0,1]$. They are the normalized values of matrices X, Y elements. In this article four transformation (normalization) methods $(X,Y)\to(P,Q)$ are applied:

$$\begin{aligned}
&\text{Max method:} && p^{(j)}_i=\frac{x^{(j)}_i}{\max\limits_{j\in\{1,2,\ldots,m\}}x^{(j)}_i}, q^{(j)}_i=\frac{y^{(j)}_i}{\max\limits_{j\in\{1,2,\ldots,m\}}y^{(j)}_i},\\
&\text{Sum method:} && p^{(j)}_i=\frac{x^{(j)}_i}{\sum\limits_{j=1}^{m}x^{(j)}_i}, q^{(j)}_i=\frac{y^{(j)}_i}{\sum\limits_{j=1}^{m}y^{(j)}_i},\\
&\text{MinMax method:} && p^{(j)}_i=\frac{x^{(j)}_i-\min\limits_{j\in\{1,2,\ldots,m\}}x^{(j)}_i}{\max\limits_{j\in\{1,2,\ldots,m\}}x^{(j)}_i-\min\limits_{j\in\{1,2,\ldots,m\}}x^{(j)}_i},\\
& && q^{(j)}_i=\frac{y^{(j)}_i-\min\limits_{j\in\{1,2,\ldots,m\}}y^{(j)}_i}{\max\limits_{j\in\{1,2,\ldots,m\}}y^{(j)}_i-\min\limits_{j\in\{1,2,\ldots,m\}}y^{(j)}_i},\\
&\text{Vector normalization:} && p^{(j)}_i=\frac{x^{(j)}_i}{\sqrt{\sum\limits_{j=1}^{m}\left(x^{(j)}_i\right)^2}}, q^{(j)}_i=\frac{y^{(j)}_i}{\sqrt{\sum\limits_{j=1}^{m}\left(y^{(j)}_i\right)^2}}.
\end{aligned} \tag{13}$$

Notice that formulas (13) applicable when all attributes are the benefit type of optimization direction, (i. e., the higher value is better, see, for example, [1]). Another case – cost type criteria, (i.e., the lower value is better) will be not discussed in this article.

Suppose that matrices P, Q and their concatenation – matrix $R = (P|Q) = \left(r_i^{(j)}\right)_{m\times(n_x+n_y)}$, i. e. $r_i^{(j)} = \begin{cases} p_i^{(j)}, & \text{if } i \leqslant n_x, \\ q_{i-n_x}^{(j)}, & \text{if } n_x+1 \leqslant i \leqslant n_x+n_y \end{cases}$ are obtained from randomly generated matrices X, Y by one of the four methods (13).

WEBIRA method will be compared with two MCDM methods – simple arithmetic average:

$$\text{AVRG: } S^{(j)} = \frac{1}{n_x+n_y}\left(\sum_{i=1}^{n_x} p_i^{(j)} + \sum_{i=1}^{n_y} q_i^{(j)}\right) \tag{14}$$

and EMDCW (Entropy Method for Determining the Criterion Weight, see [21]:

$$\begin{gathered} \text{EMDCW: } S^{(j)} = \sum_{i=1}^{n_x+n_y} w_i r_i^{(j)},\ w_i = \frac{1-e_i}{n_x+n_y-\sum\limits_{i=1}^{n_x+n_y} e_i}, \\ e_i = -\frac{1}{m}\sum_{j=1}^{m} \tilde{r}_i^{(j)} \cdot \ln\left(\tilde{r}_i^{(j)}\right),\ \tilde{r}_i^{(j)} = \frac{r_i^{(j)}}{\sum\limits_{j=1}^{m} r_i^{(j)}} \end{gathered} \tag{15}$$

Next, we provide the case study of formulas (14) and (15) application. The result of matrix (12) transformation using Max method:

$$\begin{pmatrix} 1.0000 & 1.0000 & 0.5000 & 0.6667 & 0.4000 & 1.0000 & 0.8333 & 0.6667 & 0.6667 \\ 0.8000 & 0.6667 & 0.6667 & 0.3333 & 0.6000 & 0.1667 & 0.3333 & 0.8333 & 0.5000 \\ 0.6000 & 0.6667 & 0.8333 & 0.1667 & 0.6000 & 0.3333 & 0.3333 & 0.8333 & 0.8333 \\ 0.6000 & 0.5000 & 1.0000 & 1.0000 & 1.0000 & 1.0000 & 0.6667 & 0.6667 & 0.6667 \\ 0.8000 & 0.5000 & 0.6667 & 0.5000 & 1.0000 & 1.0000 & 0.8333 & 1.0000 & 0.6667 \\ 0.4000 & 0.1667 & 0.8333 & 0.8333 & 0.4000 & 0.6667 & 1.0000 & 0.5000 & 0.1667 \\ 0.2000 & 0.3333 & 0.5000 & 0.3333 & 0.4000 & 0.3333 & 0.8333 & 1.0000 & 0.6667 \\ 0.8000 & 0.3333 & 0.3333 & 0.5000 & 0.2000 & 1.0000 & 0.3333 & 0.8333 & 1.0000 \end{pmatrix}$$

Weighting sums (15) of 8 alternatives obtained by EMDCW method are as follows:

$$0.7747; 0.4847; 0.5110; 0.8141; 0.7650; 0.5522; 0.4435; 0.5803.$$

Weighting sums calculated by AVRG method (14) are:

$$0.1470; 0.1069; 0.1135; 0.1550; 0.1521; 0.1084; 0.1004; 0.1164.$$

Thus, both methods assign the fourth as the best alternative and we treat it as a mistake, because the best is considered the first alternative. WEBIRA method was applied with initial weights values

$$\begin{aligned} W_x &= (0.3314, 0.1953, 0.1581, 0.1581, 0.1571), \\ W_y &= (1.0000, 0.0000, 0.0000, 0.0000). \end{aligned}$$

Weighted averages (2) calculated with these weights are

$$\begin{aligned} S_X^j &= (0.7739, 0.6476, 0.5813, 0.7698, 0.7043, 0.4914, 0.3259, 0.4933), \\ S_Y^j &= (1.0000, 0.1667, 0.3333, 1.0000, 1.0000, 0.6667, 0.3333, 1.0000). \end{aligned}$$

The initial loss (5) in this case is $s^0 = 0.8785$. After 134 iterations WEBIRA method allowed to reduce this value to $s^1 = 0.7147$ and the following weights were found:

$$\begin{aligned} W_x &= (0.4623, 0.1373, 0.1373, 0.1373, 0.1258), \\ W_y &= (0.5901, 0.4099, 0.0000, 0.0000). \end{aligned}$$

Weighted averages are:

$$S_X^j = (0.8100, 0.6741, 0.5816, 0.7464, 0.7245, 0.4869, 0.3029, 0.5551),$$
$$S_Y^j = (0.9316, 0.2349, 0.3333, 0.8633, 0.9316, 0.8032, 0.5382, 0.7267).$$

Ranks of alternatives according to the X: 1, 4, 5, 2, 3, 8, 6, 7 and according to the Y: 1, 5, 4, 6, 8, 7, 3, 2. So, WEBIRA method set as the best the first alternative and we treat this as the right decision. Notice that all three methods set the same three best alternatives: 1, 4 and 5.

Submit normalized matrices, calculated by other methods. MinMax method:

$$\begin{pmatrix} 1.0000 & 1.0000 & 0.2500 & 0.6000 & 0.2500 & 1.0000 & 0.7500 & 0.3333 & 0.6000 \\ 0.7500 & 0.6000 & 0.5000 & 0.2000 & 0.5000 & 0.0000 & 0.0000 & 0.6667 & 0.4000 \\ 0.5000 & 0.6000 & 0.7500 & 0.0000 & 0.5000 & 0.2000 & 0.0000 & 0.6667 & 0.8000 \\ 0.5000 & 0.4000 & 1.0000 & 1.0000 & 1.0000 & 1.0000 & 0.5000 & 0.3333 & 0.6000 \\ 0.7500 & 0.4000 & 0.5000 & 0.4000 & 1.0000 & 1.0000 & 0.7500 & 1.0000 & 0.6000 \\ 0.2500 & 0.0000 & 0.7500 & 0.8000 & 0.2500 & 0.6000 & 1.0000 & 0.0000 & 0.0000 \\ 0.0000 & 0.2000 & 0.2500 & 0.2000 & 0.2500 & 0.2000 & 0.7500 & 1.0000 & 0.6000 \\ 0.7500 & 0.2000 & 0.0000 & 0.4000 & 0.0000 & 1.0000 & 0.0000 & 0.6667 & 1.0000 \end{pmatrix}.$$

EMDCW and AVRG methods determined as the best the fifth alternative, WEBIRA – the first alternative.

Matrix transformated by Sum method:

$$\begin{pmatrix} 0.1923 & 0.2400 & 0.0937 & 0.1538 & 0.0869 & 0.1818 & 0.1612 & 0.1052 & 0.1290 \\ 0.1538 & 0.1600 & 0.1250 & 0.0769 & 0.1304 & 0.0303 & 0.0645 & 0.1315 & 0.0967 \\ 0.1154 & 0.1600 & 0.1562 & 0.0384 & 0.1304 & 0.0606 & 0.0645 & 0.1315 & 0.1612 \\ 0.1154 & 0.1200 & 0.1875 & 0.2307 & 0.2173 & 0.1818 & 0.1290 & 0.1052 & 0.1290 \\ 0.1538 & 0.1200 & 0.1250 & 0.1153 & 0.2173 & 0.1818 & 0.1612 & 0.1578 & 0.1290 \\ 0.0769 & 0.0400 & 0.1562 & 0.1923 & 0.0869 & 0.1212 & 0.1935 & 0.0789 & 0.0322 \\ 0.0384 & 0.0800 & 0.0937 & 0.0769 & 0.0869 & 0.0606 & 0.1612 & 0.1578 & 0.1290 \\ 0.1538 & 0.0800 & 0.0625 & 0.1153 & 0.0434 & 0.1818 & 0.0645 & 0.1315 & 0.1935 \end{pmatrix}.$$

In this case, as in another – vector normalization method the obtained results coincide with the Max method, i. e. EMDCW and AVRG methods set as the best the fourth alternative, while WEBIRA – the first.

Matrix transformated by Vector normalization:

$$\begin{pmatrix} 0.5103 & 0.6155 & 0.2535 & 0.3922 & 0.2222 & 0.4615 & 0.4240 & 0.2917 & 0.3442 \\ 0.4082 & 0.4103 & 0.3380 & 0.1961 & 0.3333 & 0.0769 & 0.1696 & 0.3646 & 0.2581 \\ 0.3061 & 0.4103 & 0.4225 & 0.0980 & 0.3333 & 0.1538 & 0.1696 & 0.3646 & 0.4303 \\ 0.3061 & 0.3077 & 0.5070 & 0.5883 & 0.5556 & 0.4615 & 0.3392 & 0.2917 & 0.3442 \\ 0.4082 & 0.3077 & 0.3380 & 0.2941 & 0.5556 & 0.4615 & 0.4240 & 0.4375 & 0.3442 \\ 0.2041 & 0.1025 & 0.4225 & 0.4902 & 0.2222 & 0.3076 & 0.5089 & 0.2187 & 0.0860 \\ 0.1020 & 0.2051 & 0.2535 & 0.1961 & 0.2222 & 0.1538 & 0.4240 & 0.4375 & 0.3442 \\ 0.4082 & 0.2051 & 0.1690 & 0.2941 & 0.1111 & 0.4615 & 0.1696 & 0.3646 & 0.5163 \end{pmatrix}.$$

5 Numerical experiments

Numerical experiments were conducted as follows. Random matrices X, Y generated in such a way that on average more often the best alternative is the first. Matrices P, Q are calculated by

four normalization methods and the best alternative is determined by three methods EMDCW, AVRG and WEBIRA. When the best is the first alternative the result of the experiment is marked with (+) and recorded to the table. If the best is any other (not the first) alternative (−) is recorded to the table. It is possible that WEBIRA can not set the best alternative. We then record (n). Notice, that in the cases of EMDCW and AVRG methods such experimental result is impossible. In the Table 1 the results of 5 experiments are presented. Methods EMDCW, AVRG, WEBIRA are denoted respectively as (E), (A), (W).

Table 1: Fragment of experimental results.

	Max method			MinMax method			Sum method			Vector normalization		
Nr.	(E)	(A)	(W)	(E)	(A)	(W)	(E)	(A)	(W)	(E)	(A)	(W)
1	+	+	+	-	+	+	+	+	+	+	+	+
2	+	+	+	+	+	+	+	+	n	+	+	n
3	-	-	+	+	-	+	-	-	+	-	-	+
4	-	+	+	-	+	+	-	+	+	-	+	+
5	-	-	n	-	-	n	-	-	n	-	-	n

After a series of experiments, we calculate the number of pluses (+) denoted as p in each of the 12 columns of the Table 1, the number of minuses (-) denoted as m and undetected cases (n). WEBIRA method peculiarity compared to AVRG and EMDCW – possible non zero values of parameter n. It means that WEBIRA quite often eliminates cases when it can not detect the best alternative. Consider the following indicator to compare methods performance:

$$E_n = \frac{p - m}{p + m + n}. \tag{16}$$

Indicator E_n shows reliability of the correspondent method. Our purpose is the detection of significantly different average values of E_n in the groups.

100 series of Monte Carlo experiments were carried out by 100 in each series. The common number of experiments was 10000. Random matrices estimates generation procedure is described in Chapter 3. The number of alternatives varied $m = 3, 4, ..., 50$. Table 2 presents the average values of E_n dependence on the number of alternatives m, MCDM and data normalization methods.

Conclusions and future research

WEBIRA method allows quite effectively separate the cases when it is not possible to select the best alternative. Otherwise, when the method is applicable its efficiency is significantly higher compared to the two selected simple methods: AVRG and EMDCW. This is true for all four matrices normalization methods: Max, MinMax, Sum and Vector normalization. Efficiency decreases with increasing number of alternatives, but it is still applicable for WEBIRA method, while application of AVRG and EMDCW is impossible as the number of alternatives is greater than 11, as their efficiency became negative value.

In the Figure 1 graphs average values of efficiency indicator E_n depending on the MCDM method are presented. For the Max normalization and $m = 1, 2, \ldots, 30$ the average efficiency of WEBIRA is significantly higher than efficiency of EMDCW and AVRG methods. For MinMax, Sum and Vector normalizations all 3 MCDM methods: EMDCW, AVRG and WEBIRA mutually significantly differ comparing the average values of efficiency indicator E_n when $m = 1, 2, \ldots, 15$, while WEBIRA is significantly efficient than EMDCW and AVRG for all considered numbers of alternatives.

Table 2: Average values of E_n dependence on the number of alternatives m, MCDM and data normalization methods.

	Max method			MinMax method			Sum method			Vctr method		
m	(E)	(A)	(W)	(E)	(A)	(W)	(E)	(A)	(W)	(E)	(A)	(W)
3	0.7224	0.7268	0.8837	0.4736	0.6086	0.8398	0.6944	0.7344	0.8820	0.7018	0.7286	0.8820
4	0.5762	0.5896	0.8122	0.3298	0.4868	0.7781	0.5358	0.5934	0.8006	0.5468	0.5890	0.8039
5	0.4760	0.4860	0.7721	0.2746	0.4082	0.7479	0.4288	0.4842	0.7591	0.4438	0.4770	0.7599
6	0.3870	0.4010	0.7354	0.2220	0.3358	0.7059	0.3426	0.3958	0.7075	0.3552	0.3918	0.7055
7	0.2968	0.3308	0.6694	0.1580	0.2758	0.6539	0.2556	0.3180	0.6449	0.2682	0.3200	0.6472
8	0.2060	0.2236	0.6318	0.0824	0.1744	0.6074	0.1606	0.2166	0.6003	0.1782	0.2110	0.6016
9	0.1520	0.1816	0.5999	0.0626	0.1470	0.5797	0.1134	0.1732	0.5637	0.1272	0.1724	0.5662
10	0.0998	0.1200	0.5644	0.0300	0.0884	0.5430	0.0574	0.1064	0.5314	0.0718	0.1082	0.5300
11	0.0404	0.0628	0.5275	-0.0412	0.0414	0.5149	0.0108	0.0494	0.4906	0.0168	0.0500	0.4905
13	-0.0646	-0.0374	0.4745	-0.0992	-0.0556	0.4653	-0.0776	-0.0490	0.4282	-0.0724	-0.0556	0.4274
15	-0.1356	-0.1032	0.4349	-0.1602	-0.1016	0.4293	-0.1572	-0.1208	0.3943	-0.1534	-0.1220	0.3903
20	-0.2824	-0.2512	0.3427	-0.2988	-0.2424	0.3378	-0.2930	-0.2786	0.2990	-0.2864	-0.2826	0.2989
30	-0.4726	-0.4368	0.2639	-0.4840	-0.4272	0.2606	-0.4818	-0.4660	0.2233	-0.4764	-0.4678	0.2183
50	-0.6592	-0.6250	0.1658	-0.6572	-0.6172	0.1725	-0.6604	-0.6530	0.1482	-0.6616	-0.6524	0.1488

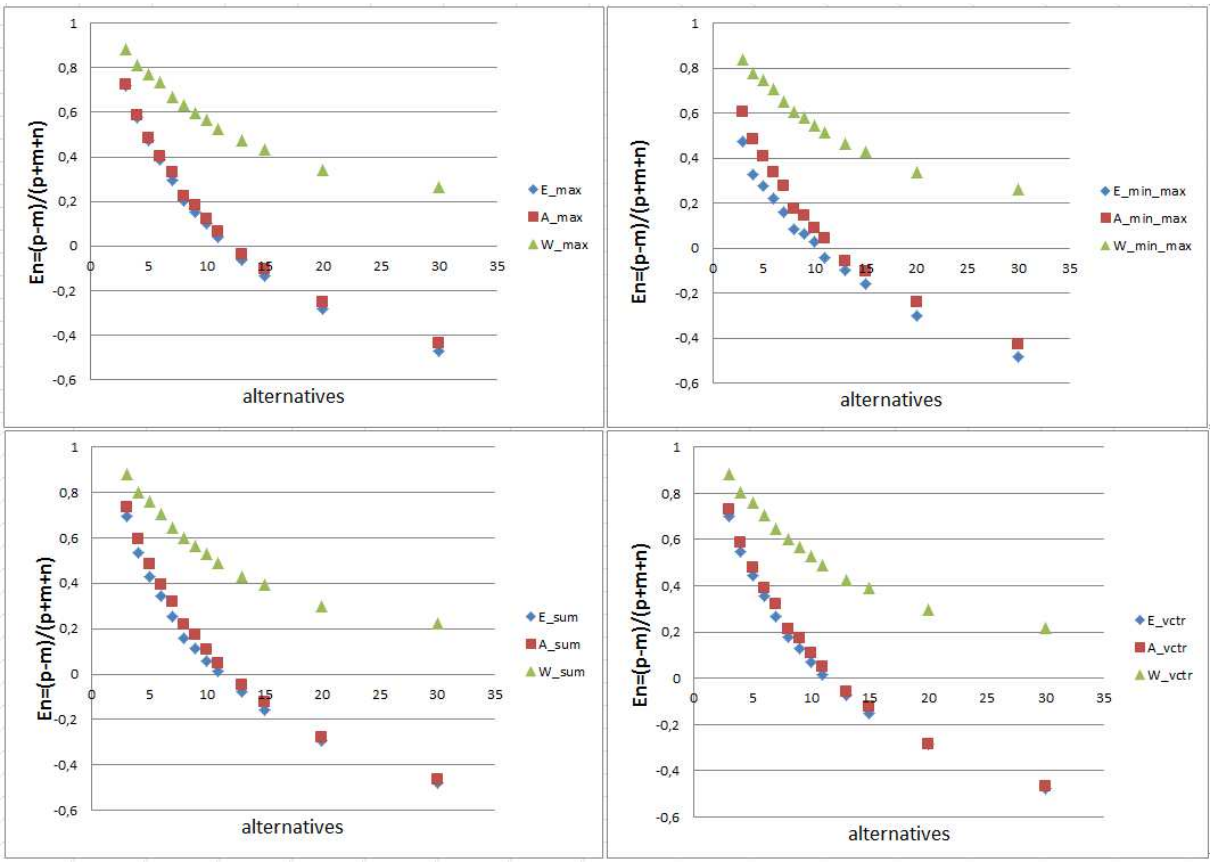


Figure 1: The dependence of E_n average values on the number of alternatives and MCDM method for Max, MinMax, Sum and Vector normalization respectively.

In the Figure 2 graphs of indicator E_n average values depending on the data normalization method are presented. WEBIRA is the least affected by the data normalization method, while EMDCW is the most dependent on normalization. One Way ANOVA was performed to compare

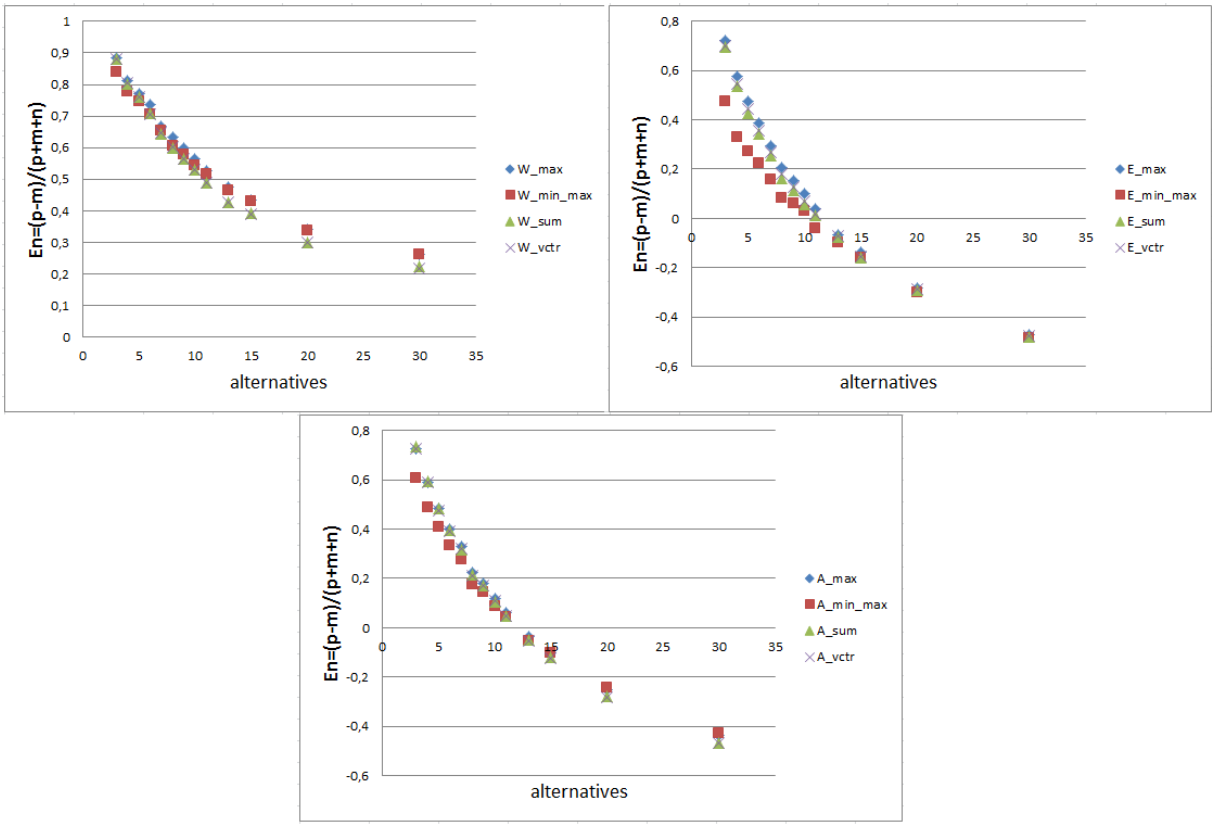


Figure 2: The dependence of E_n average values on the number of alternatives and normalization method for WEBIRA (W), EMDCW (E) and AVRG (A) method respectively.

average values of indicator E_n for the fixed m values at significance level 0.05. In the case of EMDCW (E) method it was established that E_n average values for MinMax data normalization are significantly lower than the average values obtained by Max, Sum and Vector methods when $m \leqslant 11$, and that Max normalization is significantly more efficient than MinMax and Sum methods. Hence, MinMax normalization reduces the efficiency of the EMDCW method.
In the case of AVRG (A) method average value of E_n obtained by MinMax data normalization are significantly lower than E_n averages received by the three other data normalization methods. In the case of WEBIRA (W) method at a low number of alternatives Max data normalization significantly increases the average efficiency of the method.

Random matrices generated in the article are simulating repeated expert evaluations of the same alternatives and depend on *a priori* probabilities (10)–(11). These probabilities can be such that the recognition of the best alternative will be very easy or almost impossible. It is obvious that in the first case application of WEBIRA method is irrational, and in the second – any method will not determine the best alternative. In these cases the alternatives separation requires further research. This article is limited to the case when the task of the best alternative recognition is of medium difficulty.

References describe more matrices transformation methods such as Max, MinMax, Sum and Vector normalization. Their efficiency could depend on the expert evaluation scales. In this article all attributes were assessed in 6-point scale, i. e. $x_i^{(j)}, y_i^{(j)} \in \{1, 2, 3, 4, 5, 6\}$.

There are many other simple alternative MCDM methods similar to AVRG and EMDCW. Their efficiency comparison using indicator E_n proposed in the article is a separate interesting

task. It is appropriate to look for the most efficient methods and investigate the cases when it makes sense to apply the method WEBIRA.

WEBIRA method is extended and applicable in the case of three or more subgroups of evaluating criteria. The first direction of our next research is to elaborate WEBIRA metodology for solution of practical problems where several groups of criteria naturally arise. For example, for solving sustainable management tasks where several interconnected domains such as ecology, economics, politics and environment are considered. Our other research area is the comparison of the proposed WEBIRA method with other existing methods used for solving this type of problems, i.e. when there are several natural groups of evaluation criteria. The third task is to prepare software for practical MADM problems solving by applying WEBIRA approach.

Bibliography

[1] Stanujkic,D.; Dordevic, B.; Dordevic, M. (2013); Comparative analysis of some prominent MCDM methods: A case of ranking Serbian banks, *Serbian Journal of Management*, 8(2): 213–241.

[2] Zavadskas, E.K.; Turskis, Z. (2008); A New Logarithmic Normalization Method in Games Theory, *Informatica*, 19(2): 303–314.

[3] Churchman, C.W.; Ackoff, R. (1954); An Approximate Measure of Value. *Journal of the Operations Research Society of America*, 2(2):172–187.

[4] Fishburn, P.C. (1967); Additive Utilities with Incomplete Product Sets: Application to Priorities and Assignments, *Operations Research*, 15(3):537–542.

[5] Zionts, S.; Wallenius, J. (1976); An Interactive Programming Method for Solving the Multiple Criteria Problem, *Management Science*, 26:652–663.

[6] Triantaphyllou, E. (2000); *Multi-Criteria Decision Making Methods: A Comparative Study*, Norwell, MA:Kluwer.

[7] Stanujkic, D.; Zavadskas, E.K. (2015); A modified Weighted Sum method based on the decision-maker's preferred levels of performances, *Studies in Informatics and Control*, 24(4): 461–470.

[8] Alanazi, H.O.; Abdullah, A.H.; Larbani, M. (2013); Dynamic weighted sum multi-criteria decision making: mathematical model. *International Journal of Mathematics and Statistics Invention*, 1(2): 16–18.

[9] Hwang, C.L.; Yoon, K. (1981); *Multiple Attribute Decision Making – Methods and Applications: A State-of-the-Art Survey*, Lecture Notes in Economics and Mathematical Systems, Springer, New York.

[10] Triantaphyllou, E.; Baig, K. (2005); The impact of aggregating benefit and cost criteria in four MCDA methods, *IEEE Transactions on Engineering Management*, 52(2):213–226.

[11] Kaliszewski, I.; Podkopaev, D. (2016); Simple additive weighting - A metamodel for multiple criteria decision analysis methods, *Expert Systems with Applications*, 54:155–161.

[12] Wang, Y.J. (2015); A fuzzy multi-criteria decision-making model based on simple additive weighting method and relative preference relation, *Applied Soft Computing*, 30:412–420.

[13] Tzeng, G.H.; Huang J. (2011); *Multiple Attribute Decision Making: Methods and Applications*, CRCPress, Taylor & Francis group.

[14] Mardani, A.; Jusoh, A.; Nor, MD K.; Khalifah, Z.; Zakwan, N.; Valipour, A. (2015); Multiple criteria decision-making techniques and their applications - a review of the literature from 2000 to 2014, *Economic Research–Ekonomska Istraživanja*, 28(1):516–571.

[15] MacCrimmon, K.R. (1968); *Decisionmaking Among Multiple-Attribute Alternatives: A Survey and Consolidated Approach*, Santa Monica, CA: RAND Corporation.

[16] Wang, P.; Zhu Z.Q.; Wang, Y.G. (2016); A novel hybrid MCDM model combining the SAW, TOPSIS and GRA methods based on experimental design, *Information Sciences*, 345:27–45.

[17] Ghorabaee, M.K.; Zavadskas, E.K.; Turskis, Z.; Antucheviciene, J. (2016); A new combinative distance-based assessment (CODAS) method for multi-criteria decision-making, *Economic Computation & Economic Cybernetics Studies & Research*, 50(3):25–44.

[18] Chakraborty, S.; Zavadskas, E.K. (2014); Applications of WASPAS Method in Manufacturing Decision Making, *Informatica*, 25(1):1–20.

[19] Zavadskas, E.K.; Antucheviciene, J.; Razavi Hajiagha, S.H; Hashemi, S.S. (2014); Extension of weighted aggregated sum product assessment with interval-valued intuitionistic fuzzy numbers (WASPAS-IVIF), *Applied Soft Computing*, 24:1013–1021.

[20] Peng, Y.; Zhang, Y.; Kou, G.; Shi, Y. (2012); A Multicriteria Decision Making Approach for Estimating the Number of Clusters in a Data Set, *PLoS ONE*, 7(7): e41713.

[21] Zavadskas, E.K.; Podvezko, V. (2016); Integrated Determination of Objective Criteria Weights in MCDM, *International Journal of Information Technology & Decision Making*, 15(2):267–283.

[22] Krylovas, A.; Zavadskas, E.K.; Kosareva, N.; Dadelo, S. (2014); New KEMIRA Method for Determining Criteria Priority and Weights in Solving MCDM Problem, *International Journal of Information Technology & Decision Making*, 13(1):1119–1133.

[23] Krylovas, A.; Kosareva, N.; Zavadskas, E.K. (2016); Statistical analysis of KEMIRA type weights balancing methods, *Romanian Journal of Economic Forecasting*, 19(3):19–39.

[24] Krylovas, A.; Zavadskas, E.K.; Kosareva, N. (2016); Multiple criteria decision-making KEMIRA-M method for solution of location alternatives, *Economic Research-Ekonomska Istraživanja*, 29(1): 50–65.

[25] Kosareva, N.; Zavadskas, E.K.; Krylovas, A.; Dadelo, S. (2016); Personnel ranking and selection problem solution by application of KEMIRA method, *International Journal of Computers Communications & Control*, 11(1):51–66.

[26] Zanakis, S.H.; Solomon, A.; Wishart, N.; Dublish, S. (1998); Multi-attribute decision making: A simulation comparison of select methods, *European Journal of Operational Research*, 107(3):507–529.

[27] Podvezko, V. (2011); The Comparative Analysis of MCDA Methods SAW and COPRAS, *Inzinerine Ekonomika-Engineering Economics*, 22(2):134–146.

[28] Mulliner, E.; Malys, N.; Maliene, V. (2016); Comparative analysis of MCDM methods for the assessment of sustainable housing affordability, *Omega*, 59(B):146–156.

[29] Karami, A.; Johansson, R. (2014); Utilization of Multi Attribute Decision Making Techniques to Integrate Automatic and Manual Ranking of Options, *Journal of Information Science and Engineering*, 30(2):519–534.

[30] Kolios, A.; Mytilinou, V.; Lozano-Minguez, E.; Salonitis, K. (2016); A Comparative Study of Multiple-Criteria Decision-Making Methods under Stochastic Inputs, *Energies*, 9(7):566–587.

[31] Shannon, C.E. (1948); A Mathematical Theory of Communication, *The Bell System Technical J*, 27:379–423.

[32] Čereska, A.; Podvezko, V.; Zavadskas, E.K. (2016); Operating Characteristics Analysis of Rotor Systems Using MCDM Methods, *Studies in Informatics and Control*, 25(1):59–68.

[33] Jahan, A.; Edwards, K.L. (2015); A state-of-the-art survey on the influence of normalization techniques in ranking: Improving the materials selection process in engineering design, *Materials & Design*, 65:335–342.

[34] Kaplinski, O.; Tamošaitienė, J. (2015); Analysis of Normalization Methods Influencing Results: A Review to Honour Professor Friedel Peldschus on the Occasion of his 75th Birthday, *Procedia Engineering*, 122:2–10.

INTERNATIONAL JOURNAL OF COMPUTERS COMMUNICATIONS & CONTROL
ISSN 1841-9836, 12(2):254-264, April 2017.

A Multiple Attribute Group Decision Making Method Based on 2-D Uncertain Linguistic Weighted Heronian Mean Aggregation Operator

W.H Liu, H.B. Liu, L.L. Li

WeiHua Liu*
School of Management Science and Engineering
Shandong University of Finance and Economics, Jinan 250014, China
the Second Ring Road No. 7366
Lixia District of Jinan, Shandong Province,China
*Corresponding author: WeihuaLiu@sdufe.edu.cn

HaiBo Liu
School of Chemistry and Chemical Engineering
QuFu Normal University, QuFu 273165, China
qflhb@163.com

LingLing Li
School of Management Science and Engineering
Shandong University of Finance and Economics, Jinan 250014, China
Lilingling@sdufe.edu.cn

Abstract: 2-Dimension uncertain linguistic variables can describe both subjective evaluation result of attributes and reliability of the evaluation results in multiple attribute decision making problems. However, it is difficult to aggregate these evaluation information and give comprehensive results. Heronian mean (HM) has the characteristic of capturing the correlations between aggregated arguments and is extended to solve this problem. The 2-dimension uncertain linguistic weighted HM aggregation(2DULWHMA) operator is employed in this paper. Firstly, the definition, properties, expectations and the operational laws the 2-dimension uncertain linguistic variables are investigated. Furthermore, the properties of the 2DULWHMA operators, such as commutativity, idempotency and monotonicity, etc. are studied. Some special cases of the generalized parameters in these operators are analyzed. Finally, an example is given to demonstrate the effectiveness and feasibility of the proposed method.
Keywords: 2-Dimension (2-D) uncertain linguistic variables; Heronian mean; aggregation operator.

1 Introduction

Multiple attribute decision making (MADM) refers to ranking and selecting the best alternatives by utilizing the known information. It has been widely employed in economic, science and technology etc. Since Churchman et al. [1] introduced the multi-attribute decision making and employed it in enterprise investment decisions, the classical multi-attribute decision making has attracted more and more research attention. However, owing to the complexity and uncertainty of objective things and the fuzziness of human thought, a large number of MADM problems are uncertain, which are called uncertain multiple attribute decision making problems. Those uncertain multiple attribute decision making problems are difficult to evaluate alternatives for decision makers using real numbers in many cases [25].It is more reasonable and natural expressed by combining linguistic information (good, fair, poor) with fuzzy term (slightly, very, mightily,

Copyright © 2006-2017 by CCC Publications

extremely, obviously) . The researches on the MADM based on linguistic variables have made the fruitful achievements[614].

However, there exist a kind of the linguistic fuzzy MADM problems in practical decision-making, such as review of the science technology project, blind evaluate of economic industry system, etc. In these decision-making problems, decision makers not only assess all the indicators of evaluation objects, but also estimate the familiarity with the given results. Therefore, Zhu et al. [15] presented the concept of 2-dimension linguistic assessment information to solve this kind of decision making problems. Liu et al. [16]extended 2-dimension linguistic information to deal with the multiple attribute group decision making problems with unknown weight. 2-dimension linguistic employs two class linguistic information to describe the judgment on the object representing evaluation result and reliability of evaluation respectively. This can easily distinguish indetermination between decision making problems and subjective understanding, which is helpful to express opinions more accurately for decision makers. When 2-dimension linguistic assessment information is described by uncertain linguistic variables, it is called as 2-dimension uncertain linguistic variables.

At present, research based on 2-dimension linguistic information mainly focus on the following. Aggregation operators is a hot topic. 2-dimension uncertain linguistic power generalized weighted aggregation operator is proposed and some properties are discussed[15]. 2-dimension uncertain linguistic generalized hybrid weighted average operator is proposed and employed in technological innovation ability evaluation[17]. Liu and Qi[18] proposed some generalized dependent aggregation operators for 2-dimension linguistic information and developed a group decision making method based on these operators. 2-dimension uncertain linguistic density geometric aggregation operators and 2-dimension uncertain linguistic density generalized aggregation operators is proposed and used in an example[19]. Another topic is operational rules, the operational rules of 2-dimension uncertain linguistic variables are improved by transferring it into intuitionistic fuzzy numbers to make the operations more accurate[20]. Yu et al. [21] transformed 2-dimension linguistic information into generalized triangular fuzzy number, and proposed 2-dimension linguistic weighted average (2DLWA) operator and 2-dimension linguistic ordered weighted average (2DLOWA) operator.

Although the 2-dimension linguistic variables can reflect the evaluation of decision makers on objects, some information to be aggregated can still be omitted easily, which makes the integrated result cannot be better to express alternatives synthetically. However Heronian mean operator can capture the correlations of the aggregated arguments, so we combine 2-dimension linguistic variables with Heronian mean and propose some 2-dimension uncertain linguistic Heronian mean aggregation operators and weighted Heronian mean aggregation operators, further apply them to solve the MADM problems.

In order to do so, the remainder of this paper is shown as follows. In Sect. 2, we briefly introduce the operational rules of 2-dimension linguistic variables and the Heronian mean. In Sect. 3, some Heronian mean aggregation operators based on 2-dimension linguistic variables are proposed and commutativity, idempotency and monotonicity are studied.In Sect.4 detailed calculating steps are given to solve the group decision making problem with 2-dimension uncertain linguistic information. In Sect. 5, we use an illustrate example to verify the efficiency of the proposed method, and some special cases are also discussed. In the last section, the conclusions are given.

2 2-Dimension uncertain linguistic variable and Heronian mean

2.1 Uncertain linguistic variable

When decision makers need to express the qualitative information, generally it is necessary to set an appropriate linguistic assessment set in advance. Let the linguistic assessment set be $S = \{s_l \,|l = 1, 2, \cdots, L-1\}$, where s_l represents a linguistic variable, L is odd number. when $L = 7$, it is represented as = (extremely poor, very poor, poor, medium, good, very good, extremely good).

2.2 2-Dimension uncertain linguistic variable

In many cases, decision makers should give both evaluation conclusion of evaluation objects and other similar the reliability of their evaluation. In order to reliably describe decision makers judgment to the evaluated object, 2-dimension uncertain linguistic variable $\tilde{s} = (s_F, s_T)$ has been proposed[22]. s_F= $\{\dot{s}_i \,|i = 1, 2, \cdots, n-1\}$represents decision makers judgment to an evaluated object and s_T= $\{\ddot{s}_i \,|i = 1, 2, \cdots, m-1\}$ represents the subjective evaluation on the reliability of their given results.

Definition 1. Let $\tilde{s} = ([\dot{s}_a, \dot{s}_b], [\ddot{s}_c, \ddot{s}_d])$ where $\dot{s}_a, \dot{s}_b \in s_F$ is I class uncertain linguistic information, $\ddot{s}_c, \ddot{s}_d \in s_T$ is II class uncertain linguistic information [21]. In order to minimize the loss of linguistic information, the discrete linguistic assessment sets of 2-dimension uncertain linguistic information are extended to continuous linguistic assessment sets, such that $\dot{s}_a, \dot{s}_b \in s_F = \{s_i \,|i \in [0, t]\}$ and Let $\ddot{s}_c, \ddot{s}_d \in s_T = \{s_j \,|j \in [0, t]\}$ be the set of all 2-dimension uncertain linguistic variables.

Consider any three uncertain linguistic variables $\tilde{s} = ([\dot{s}_a, \dot{s}_b][\ddot{s}_c, \ddot{s}_d])$,$\tilde{s}_i = ([\dot{s}_{a_i}, \dot{s}_{b_i}][\ddot{s}_{c_i}, \ddot{s}_{d_i}])$ and $\tilde{s}_j = ([\dot{s}_{a_j}, \dot{s}_{b_j}][\ddot{s}_{c_j}, \ddot{s}_{d_j}])$.let $\lambda \geq 0$,then the operational rules are defined as follows[15][23˜24]:

$$\tilde{s}_i \oplus \tilde{s}_j = ([\dot{s}_{a_i}, \dot{s}_{b_i}][\ddot{s}_{c_i}, \ddot{s}_{d_i}]) \oplus ([\dot{s}_{a_j}, \dot{s}_{b_j}][\ddot{s}_{c_j}, \ddot{s}_{d_j}]) = ([\dot{s}_{a_i+a_j}, \dot{s}_{b_i+b_j}][\ddot{s}_{\min(c_i,c_j)}, \ddot{s}_{\min(d_i,d_j)}]) \quad (1)$$

$$\tilde{s}_i \otimes \tilde{s}_j = ([\dot{s}_{a_i}, \dot{s}_{b_i}][\ddot{s}_{c_i}, \ddot{s}_{d_i}]) \otimes ([\dot{s}_{a_j}, \dot{s}_{b_j}][\ddot{s}_{c_j}, \ddot{s}_{d_j}]) = ([\dot{s}_{a_i \times a_j}, \dot{s}_{b_i \times b_j}][\ddot{s}_{\min(c_i,c_j)}, \ddot{s}_{\min(d_i,d_j)}]) \quad (2)$$

$$\lambda\tilde{s} = \lambda([\dot{s}_a, \dot{s}_b][\ddot{s}_c, \ddot{s}_d]) = ([\dot{s}_{\lambda a}, \dot{s}_{\lambda b}][\ddot{s}_c, \ddot{s}_d]) \quad (3)$$

$$\tilde{s}^{\lambda} = ([\dot{s}_a, \dot{s}_b][\ddot{s}_c, \ddot{s}_d])^{\lambda} = ([\dot{s}_{a^{\lambda}}, \dot{s}_{b^{\lambda}}][\ddot{s}_c, \ddot{s}_d]) \quad (4)$$

According to above equation from (1) to (4),following relationship can be easily proved:

$$\tilde{s}_i \oplus \tilde{s}_j = \tilde{s}_j \oplus \tilde{s}_i \quad (5)$$

$$\tilde{s}_i \otimes \tilde{s}_j = \tilde{s}_j \otimes \tilde{s}_i \quad (6)$$

$$\lambda(\tilde{s}_i \oplus \tilde{s}_j) = \lambda\tilde{s}_i \oplus \lambda\tilde{s}_j \quad (7)$$

$$\lambda_1\tilde{s} \oplus \lambda_2\tilde{s} = (\lambda_1 + \lambda_2)\tilde{s} \quad (8)$$

$$\tilde{s}^{\lambda_1} \oplus \tilde{s}^{\lambda_2} = \tilde{s}^{(\lambda_1+\lambda_2)} \quad (9)$$

Definition 2. Let $\tilde{s} = ([\dot{s}_a, \dot{s}_b][\ddot{s}_c, \ddot{s}_d])$ be a 2-dimension uncertain linguistic variable, m is the length of s_F and n is the length of s_T ,then the expectation $E(\tilde{s})$ of $\tilde{s}$ is defined as

$$E(\tilde{s}) = \frac{a+b}{2(m-1)} \times \frac{c+d}{2(n-1)} \quad (10)$$

For any 2-dimension uncertain linguistic variable $\tilde{s}_i$ and $\tilde{s}_j$, if $E(\tilde{s}_i) \geq E(\tilde{s}_j)$ then $\tilde{s}_i \geq \tilde{s}_j$,or vice versa.

2.3 Heronian mean

Heronian Mean is one of aggregation methods, which can reflect the interrelationship of the input arguments[25][26].

Definition 3. Let $s_i(i = 1, 2, \cdots, n)$ be a collection of nonnegative numbers, then HM is defined as

$$HM(s_1, s_2, \cdots, s_n) = \frac{2}{n(n+1)} \sum_{i=1}^{n} \sum_{j=i}^{n} \sqrt{a_i a_j} \tag{11}$$

Definition 4. Let $p, q \geq 0$ and p , q do not take the value 0 simultaneously, then generalized Heronian mean(GHM) is defined

$$GHM(s_1, s_2, \cdots, s_n) = \left(\frac{2}{n(n+1)} \sum_{i=1}^{n} \sum_{j=i}^{n} a_i^p a_j^q \right)^{\frac{1}{p+q}} \tag{12}$$

3 The 2-dimension uncertain linguistic weighted Heronian mean aggregation operators

Definition 5. Let $p, q \geq 0$ and p, q do not take the value 0 simultaneously. $\tilde{s}_i = ([\dot{s}_{a_i}, \dot{s}_{b_i}][\ddot{s}_{c_i}, \ddot{s}_{d_i}])$ be a collection of 2-dimension uncertain linguistic variables. 2-dimension uncertain linguistic generalized Heronian mean aggregation (2DULGHMA) operators :$\Omega^n \to \Omega$ is defined as

$$2DULGHMA(\tilde{s}_1, \tilde{s}_2, \cdots, \tilde{s}_n) = \left(\frac{2}{n(n+1)} \sum_{i=1, j=i}^{n} \tilde{s}_i^p \tilde{s}_j^q \right)^{\frac{1}{p+q}} \tag{13}$$

According to the operational rules of 2-dimension uncertain linguistic variables, formula(13) can be transformed into the following formŁş

$$\begin{aligned} &2DULGHMA(\tilde{s}_1, \tilde{s}_2, \cdots, \tilde{s}_n) = \left(\tfrac{2}{n(n+1)} \sum_{i=1, j=i}^{n} \tilde{s}_i^p \tilde{s}_j^q \right)^{\frac{1}{p+q}} \\ &= \left(\left[\dot{s}_{(\frac{2}{n(n+1)} \sum_{i=1, j=i}^{n} a_i^p a_j^q)^{\frac{1}{p+q}}}, \dot{s}_{(\frac{2}{n(n+1)} \sum_{i=1, j=i}^{n} b_i^p b_j^q)^{\frac{1}{p+q}}} \right], \left[\min_i \ddot{s}_{c_i}, \min_i \ddot{s}_{d_i} \right] \right) \end{aligned} \tag{14}$$

Theorem 6. *(Commutativity). Let* $(\tilde{s}'_1, \tilde{s}'_2, \cdots, \tilde{s}'_n)$ *is any permutation of* $(\tilde{s}_1, \tilde{s}_2, \cdots, \tilde{s}_n)$ *then* $2DULGHMA(\tilde{s}_1, \tilde{s}_2, \cdots, \tilde{s}_n) = 2DULGHMA(\tilde{s}'_1, \tilde{s}'_2, \cdots, \tilde{s}'_n)$

Proof: Since $(\tilde{s}'_1, \tilde{s}'_2, \cdots, \tilde{s}'_n)$ is any permutation of $(\tilde{s}_1, \tilde{s}_2, \cdots, \tilde{s}_n)$,then

$$\begin{aligned} &2DULGHMA(\tilde{s}_1, \tilde{s}_2, \cdots, \tilde{s}_n) = \left(\tfrac{2}{n(n+1)} \sum_{i=1, j=i}^{n} \tilde{s}_i^p \tilde{s}_j^q \right)^{\frac{1}{p+q}} \\ &= \left(\tfrac{2}{n(n+1)} \sum_{i=1, j=i}^{n} \tilde{s}_i'^p \tilde{s}_j'^q \right)^{\frac{1}{p+q}} = 2DULGHMA(\tilde{s}'_1, \tilde{s}'_2, \cdots, \tilde{s}'_n) \end{aligned} \tag{15}$$

□

Theorem 7. *(Idempotency). Let* $\tilde{s}_i = ([\dot{s}_{a_i}, \dot{s}_{b_i}][\ddot{s}_{c_i}, \ddot{s}_{d_i}])(i = 1, 2, \cdots, n)$ *be a collection of uncertainty linguistic variables. If all* $\tilde{s}_i$ *are equal, for all* i *,*$\tilde{s}_i = \tilde{s} = ([\dot{s}_a, \dot{s}_b][\ddot{s}_c, \ddot{s}_d])$ *then* $2DULGHMA(\tilde{s}_1, \tilde{s}_2, \cdots, \tilde{s}_n) = \tilde{s}$

Proof:

$$\begin{aligned} &2DULGHMA(\tilde{s}_1, \tilde{s}_2, \cdots, \tilde{s}_n) = 2DULGHMA(\tilde{s}, \tilde{s}, \cdots, \tilde{s}) \\ &= \left(\frac{2}{n(n+1)} \sum_{i=1,j=i}^{n} \tilde{s}^p \tilde{s}^q \right)^{\frac{1}{p+q}} = \left(\frac{2}{n(n+1)} (n + (n-1) + \cdots 1) \tilde{s}^{p+q} \right)^{\frac{1}{p+q}} \\ &= \tilde{s} \end{aligned} \tag{16}$$

□

Theorem 8. *(Monotonicity). Let* $\tilde{s}_i = ([\dot{s}_{a_i}, \dot{s}_{b_i}][\ddot{s}_{c_i}, \ddot{s}_{d_i}])$ *and* $\tilde{s}'_i = ([\dot{s}'_{a_i}, \dot{s}'_{b_i}][\ddot{s}'_{c_i}, \ddot{s}'_{d_i}])(i = 1, 2, \cdots, n)$ *be a collection of uncertainty linguistic variables. If all* $\tilde{s}_i$ *are 2-dimension uncertain linguistic variables. For all* i*,if* $\tilde{s}_i \leq \tilde{s}_i{}'$ *,then* $2DULGHMA(\tilde{s}_1, \tilde{s}_2, \cdots, \tilde{s}_n) \leq 2DULGHMA(\tilde{s}'_1, \tilde{s}'_2, \cdots, \tilde{s}'_n)$

Proof:

$$\begin{aligned} &2DULGHMA(\tilde{s}_1, \tilde{s}_2, \cdots, \tilde{s}_n) \\ &= \left(\frac{2}{n(n+1)} \sum_{i=1,j=i}^{n} \tilde{s}_i^p \tilde{s}_j^q \right)^{\frac{1}{p+q}} \leq \left(\frac{2}{n(n+1)} \sum_{i=1,j=i}^{n} \tilde{s}_i{}'^p \tilde{s}_i{}'^q \right)^{\frac{1}{p+q}} \\ &= 2DULGHMA(\tilde{s}'_1, \tilde{s}'_2, \cdots, \tilde{s}'_n) \end{aligned} \tag{17}$$

□

Theorem 9. *(Boundedness).Let 2DULGHMA operator* $\tilde{s}_i = ([\dot{s}_{a_i}, \dot{s}_{b_i}][\ddot{s}_{c_i}, \ddot{s}_{d_i}])(i = 1, 2, \cdots, n)$ *,then* $\min_i(\widetilde{s}_i) \leq 2DULGHMA(\widetilde{s}_1, \widetilde{s}_2, \cdots, \widetilde{s}_n) \leq \max_i(\widetilde{s}_i)$

Proof: Let $\tilde{s}_{\min} = \min(\tilde{s}_1, \tilde{s}_2, \cdots, \tilde{s}_n)$ $\tilde{s}_{\max} = \max(\tilde{s}_1, \tilde{s}_2, \cdots, \tilde{s}_n)$,since $\tilde{s}_{\min} \leq \tilde{s}_i \leq \tilde{s}_{\max}$,then

$$\left(\frac{2}{n(n+1)} \sum_{i=1,j=i}^{n} \tilde{s}_{\min}^p \tilde{s}_{\min}^q \right)^{\frac{1}{p+q}} \leq \left(\frac{2}{n(n+1)} \sum_{i=1,j=i}^{n} \tilde{s}_i^p \tilde{s}_j^q \right)^{\frac{1}{p+q}} \leq \left(\frac{2}{n(n+1)} \sum_{i=1,j=i}^{n} \tilde{s}_{\max}^p \tilde{s}_{\max}^q \right)^{\frac{1}{p+q}} \tag{18}$$

□

Eq.(14) assumes that all of arguments being aggregated are of equal importance. However, in many real cases, the importance degrees are not equal. Thus, we need to assign different weights for different arguments, and further define a new aggregation operator to process this case.

Because unfair evaluation information may be provided by some decision makers, the weight of evaluation information are larger when they are consistent with other evaluation information given by other decision makers. Otherwise, smaller weight will be set. As result, the influence of outlier values can be reduced, and the decision results based on these operators will be more reasonable and reliable.

Definition 10. Let $p, q \geq 0$ and p, q do not take the value 0 simultaneously. $\tilde{s}_i = ([\dot{s}_{a_i}, \dot{s}_{b_i}][\ddot{s}_{c_i}, \ddot{s}_{d_i}])$ be a collection of 2-dimension uncertain linguistic variables whose weight vector is $\omega = (\omega_1, \omega_2, \cdots, \omega_n)^T$ satisfying $\omega_i \in [0, 1]$ and $\sum_{i=1}^{n} \omega_j = 1$, then 2DULWHMA is called the 2-dimension uncertain linguistic weighted HM aggregation (2DULWHMA) operator: $\Omega^n \to \Omega$.

If

$$
\begin{aligned}
&2DULWHMA(\tilde{s}_1, \tilde{s}_2, \cdots, \tilde{s}_n) \\
&= \left(\frac{2}{n(n+1)} \sum_{i=1,j=i}^{n} (\omega_i \tilde{s}_i)^p (\omega_j \tilde{s}_j)^q \right)^{\frac{1}{p+q}} \\
&= \left(\left[\dot{s}_{\left(\frac{2}{n(n+1)} \sum_{i=1,j=i}^{n} (\omega_i a_i)^p (\omega_j a_j)^q \right)^{\frac{1}{p+q}}}, \dot{s}_{\left(\frac{2}{n(n+1)} \sum_{i=1,j=i}^{n} (\omega_i b_i)^p (\omega_j b_j)^q \right)^{\frac{1}{p+q}}} \right], \left[\min_i \ddot{s}_{c_i}, \min_i \ddot{s}_{d_i} \right] \right)
\end{aligned} \tag{19}
$$

Similar to Theorems 69,we can prove that the 2DULWHMA operator with the following properties:

Theorem 11. *(Commutativity) Let* $(\tilde{s}'_1, \tilde{s}'_2, \cdots, \tilde{s}'_2)$ *is any permutation of* $(\tilde{s}_1, \tilde{s}_2, \cdots, \tilde{s}_n)$ *then* $2DULWHMA(\tilde{s}_1, \tilde{s}_2, \cdots, \tilde{s}_n) = 2DULWHMA(\tilde{s}'_1, \tilde{s}'_2, \cdots, \tilde{s}'_2)$

Theorem 12. *(Idempotency) Let* $\tilde{s}_i = \tilde{s}(i = 1, 2, \cdots, n)$ *then* $2DULWHMA(\tilde{s}_1, \tilde{s}_2, \cdots, \tilde{s}_n) = \tilde{s}$

Theorem 13. *(Monotonicity) Let* $\tilde{s}_i$ *and* $\tilde{s}'_i$ $(i = 1, 2, \cdots, n)$ *are 2-dimension uncertain linguistic variables. For all* i*, if* $\tilde{s}_i \leq \tilde{s}'_i$ *,then* $2DULWHMA(\tilde{s}_1, \tilde{s}_2, \cdots, \tilde{s}_n) \leq 2DULWHMA(\tilde{s}'_1, \tilde{s}'_2, \cdots, \tilde{s}'_n)$

Theorem 14. *(Boundedness) The operator lies between the max and min operators:*
$\min(\tilde{s}_1, \tilde{s}_2, \cdots, \tilde{s}_n) \leq 2DULWHMA(\tilde{s}_1, \tilde{s}_2, \cdots, \tilde{s}_n) \leq \max(\tilde{s}_1, \tilde{s}_2, \cdots, \tilde{s}_n)$

4 A Method for group decision making based on Weighted HM under 2-dimension uncertain linguistic environment

In this section, we consider a group decision making problem with 2-dimension uncertain linguistic information. 2DULWHMA proposed in section 3 will be used to solve a group decision making problem.

Let $A = \{A_1, A_2, \cdots, A_m\}$ be a discrete set of alternatives and $C = \{C_1, C_2, \cdots, C_n\}$ be the set of attributes, whose weighting vector is $\omega = \{\omega_1, \omega_2, \cdots, \omega_n\}^T$ such that $\omega_i \geq 0$ and $\sum_{i=1}^{n} \omega_i = 1$. Let $D = \{D_1, D, \cdots, D_t\}$ be the set of decision makers, and $\lambda = (\lambda_1, \lambda_2, \cdots, \lambda_t)$ is the expert weight vector such that $\lambda_k \geq 0$ and $\sum_{k=1}^{t} \lambda_k = 1$,The decision matrix $D_k = [\widetilde{d}_{ij}^k]_{m \times n}$ is 2-dimension uncertain linguistic variable about the attribute C_j for the alternative A_i and $\widetilde{d}_{ij}^k = ([\dot{s}_{a_{ij}^k}, \dot{s}_{a_{ij}^k}][\ddot{s}_{c_{ij}^k}, \ddot{s}_{d_{ij}^k}])$ where $\dot{s}_a, \dot{s}_b \in s_F$ and $\ddot{s}_c, \ddot{s}_d \in s_T$ is uncertain linguistic information. Then the ranking of alternatives is required. The method involves the following steps:

Step 1: Utilize the 2DULWHMA to aggregate the evaluation values of each expert $(j = 1, 2, \cdots, n)$

$$\begin{aligned}
&\widetilde{d}_{ij} = 2DULWHMA(\widetilde{d}_{ij}^{1}, \widetilde{d}_{ij}^{2}, \cdots, \widetilde{d}_{ij}^{t}) \\
&= \left(\frac{2}{t(t+1)} \sum_{k=1,l=k}^{t} (\lambda_{ij}^{k} \widetilde{d}_{ij}^{k})^{p} (\lambda_{ij}^{l} \widetilde{d}_{ij}^{l})^{q} \right)^{\frac{1}{p+q}} \\
&= \left(\left[\dot{s}_{\left(\frac{2}{t(t+1)} \sum_{k=1,l=k}^{t} (\lambda_{ij}^{k} a_{ij}^{k})^{p} (\lambda_{ij}^{l} a_{ij}^{l})^{q} \right)^{\frac{1}{p+q}}}, \dot{s}_{\left(\frac{2}{t(t+1)} \sum_{k=1,l=k}^{t} (\lambda_{ij}^{k} b_{ij}^{k})^{p} (\lambda_{ij}^{l} b_{ij}^{l})^{q} \right)^{\frac{1}{p+q}}} \right], \left[\min_{k} \ddot{s}_{c_{ik}}, \min_{k} \ddot{s}_{d_{ik}} \right] \right)
\end{aligned} \tag{20}$$

Step 2: Aggregate the evaluation information of each attribute by 2DULWHMA operator based on the following formula ($i = 1, 2, \cdots, m$)

$$\begin{aligned}
&\widetilde{d}_{i} = 2DULWHMA(\widetilde{d}_{i1}, \widetilde{d}_{i2}, \cdots, \widetilde{d}_{im}) \\
&= \left(\frac{2}{m(m+1)} \sum_{j=1,k=j}^{m} (\omega_{j} \widetilde{d}_{ij})^{p} (\omega_{k} \widetilde{d}_{ik})^{q} \right)^{\frac{1}{p+q}} \\
&= \left(\left[\dot{s}_{\left(\frac{2}{m(m+1)} \sum_{j=1,k=j}^{m} (\omega_{j} a_{ij})^{p} (\omega_{k} a_{ik})^{q} \right)^{\frac{1}{p+q}}}, \dot{s}_{\left(\frac{2}{m(m+1)} \sum_{j=1,k=j}^{n} (\omega_{j} b_{ij})^{p} (\omega_{k} b_{ik})^{q} \right)^{\frac{1}{p+q}}} \right], \left[\min_{j} \ddot{s}_{c_{ij}}, \min_{j} \ddot{s}_{d_{ij}} \right] \right)
\end{aligned} \tag{21}$$

Step 3: Calculate the expectation $E(\widetilde{d}_i)$ of 2-dimension uncertain linguistic variable according to Equation(10).

Step 4: Rank all the alternatives and select the best ones in accordance with the ranking of $E(\widetilde{s})$.

5 Examples illustration and discussion

In this section, the method are illustrated through an application case of evaluation in extra-efficient economic industry system of Shandong province. From this example, we explains the actual application and the effectiveness of the proposed method. In order to evaluate industry ecological level, four typical industries are selected and expressed by $\{a_1, a_2, a_3, a_4\}$. a_1 is for ecological agriculture,a_2 is for environmentally friendly industry, a_3 is for energy saving industry, and a_4 is for circular economy. Three experts $\{e_1, e_2, e_3\}$ were invited to evaluate these projects by following indexes expressed by $\{ c_1, c_2, c_3, c_4\}$. c_1 expressed *Industrial efficiency* reflectting the quantity and quality of industry economic growth, and it could be evaluated by the energy consumption, material consumption and water consumption, and etc. c_2 expressed *industrial structure coordination* reflecting industry coordination among the economic benefits, environmental benefits and social benefits, and it could be evaluated by industrial spatial distribution harmony, industrial structure harmony and R&D to investment industrial gross output etc. c_3 expressed *environment benefits* reflecting impacts on the industry environmental over entire product life cycle ,it could be evaluated by cleaner production, pollution control, reclamation of wastes and condition of work safety etc. c_4 expressed *performance indicator* measuring the industry development contribution for social progress ,it could be evaluate by total tax and profit payment, total assets contribution and green GDP per person etc.

Supposed that $\lambda = (0.243, 0.514, 0.243)$ is the weight vector of the five experts, and $\omega = (0.25, 0.27, 0.25, 0.23)$ is the index weight. The index values given by the experts take the form of 2- dimension uncertain linguistic variables, and they are shown in Tables 1-3.

The experts adopted I class linguistic set and the II class linguistic set.

According to step 1- step 4, we got the following result as table 4.

The 2DULWHMA were utilized to aggregate the arguments, some different overall attribute value $\widetilde{d}_i$ of the alternatives $a_i\,(i=1,2,3,4)$ were list in Table 4. From this table, we could find that the overall attribute values of each alternative depended on the choice of the parameters p and q,but the ranking was kept unchanged. According to table 4,a1 has higher ecological level.

Conclusions and future works

The multiple attribute decision making problems based on 2-dimension uncertain linguistic variables have been applied to a wide range of areas. Compared with the traditional uncertain linguistic variables, 2-dimension uncertain linguistic variables add a subjective evaluation on the reliability of the evaluation results given by decision makers, so they can better express fuzzy information.

In addition, the HM can take all the decision arguments and their relationships into account. Based on HM and 2-dimension uncertain linguistic variables, 2DULWHMA operators are developed in this paper. Some desirable properties, such as commutativity, idempotency, monotonicity and boundedness are discussed. Moreover, different p, q cannot affect ranking of alternatives , which proved that it is a flexible multiple attribute decision making method in that the decision makers, and they can choose different values of the parameters p and q according to their actual needs. Finally, to demonstrate the effectiveness and feasibility of the developed method, an example about industry ecological level is given. In further research, the proposed operators and methods can be extended to other fuzzy information.

Table 1: the index values of industry given by expert e1

industry	attribute(c1)	attribute (c2)	attribute (c3)	attribute (c4)
a1	$([\dot{s}_4,\dot{s}_5],[\ddot{s}_2,\ddot{s}_3])$	$([\dot{s}_2,\dot{s}_3],[\ddot{s}_3,\ddot{s}_3])$	$([\dot{s}_3,\dot{s}_5],[\ddot{s}_4,\ddot{s}_4])$	$([\dot{s}_4,\dot{s}_5],[\ddot{s}_1,\ddot{s}_2])$
a2	$([\dot{s}_3,\dot{s}_4],[\ddot{s}_2,\ddot{s}_3])$	$([\dot{s}_3,\dot{s}_5],[\ddot{s}_3,\ddot{s}_3])$	$([\dot{s}_4,\dot{s}_4],[\ddot{s}_4,\ddot{s}_4])$	$([\dot{s}_4,\dot{s}_4],[\ddot{s}_1,\ddot{s}_2])$
a3	$([\dot{s}_2,\dot{s}_3],[\ddot{s}_2,\ddot{s}_3])$	$([\dot{s}_3,\dot{s}_4],[\ddot{s}_3,\ddot{s}_3])$	$([\dot{s}_3,\dot{s}_4],[\ddot{s}_4,\ddot{s}_4])$	$([\dot{s}_4,\dot{s}_5],[\ddot{s}_1,\ddot{s}_2])$
a4	$([\dot{s}_5,\dot{s}_6],[\ddot{s}_3,\ddot{s}_4])$	$([\dot{s}_1,\dot{s}_2],[\ddot{s}_3,\ddot{s}_3])$	$([\dot{s}_3,\dot{s}_4],[\ddot{s}_4,\ddot{s}_4])$	$([\dot{s}_3,\dot{s}_4],[\ddot{s}_1,\ddot{s}_2])$

Table 2: the index values of industry given by expert e2

industry	attribute(c1)	attribute (c2)	attribute (c3)	attribute (c4)
a1	$([\dot{s}_4,\dot{s}_4],[\ddot{s}_3,\ddot{s}_4])$	$([\dot{s}_3,\dot{s}_4],[\ddot{s}_2,\ddot{s}_3])$	$([\dot{s}_4,\dot{s}_4],[\ddot{s}_3,\ddot{s}_3])$	$([\dot{s}_5,\dot{s}_6],[\ddot{s}_3,\ddot{s}_4])$
a2	$([\dot{s}_4,\dot{s}_5],[\ddot{s}_2,\ddot{s}_3])$	$([\dot{s}_3,\dot{s}_3],[\ddot{s}_2,\ddot{s}_3])$	$([\dot{s}_4,\dot{s}_5],[\ddot{s}_3,\ddot{s}_3])$	$([\dot{s}_4,\dot{s}_4],[\ddot{s}_3,\ddot{s}_3])$
a3	$([\dot{s}_3,\dot{s}_4],[\ddot{s}_3,\ddot{s}_3])$	$([\dot{s}_4,\dot{s}_5],[\ddot{s}_2,\ddot{s}_3])$	$([\dot{s}_2,\dot{s}_3],[\ddot{s}_3,\ddot{s}_3])$	$([\dot{s}_3,\dot{s}_4],[\ddot{s}_3,\ddot{s}_4])$
a4	$([\dot{s}_5,\dot{s}_5],[\ddot{s}_3,\ddot{s}_4])$	$([\dot{s}_4,\dot{s}_5],[\ddot{s}_2,\ddot{s}_3])$	$([\dot{s}_2,\dot{s}_2],[\ddot{s}_3,\ddot{s}_3])$	$([\dot{s}_3,\dot{s}_4],[\ddot{s}_3,\ddot{s}_4])$

Table 3: the index values of industry given by expert e3

industry	attribute(c1)	attribute (c2)	attribute (c3)	attribute (c4)
a1	$([\dot{s}_5,\dot{s}_5],[\ddot{s}_2,\ddot{s}_3])$	$([\dot{s}_3,\dot{s}_3],[\ddot{s}_2,\ddot{s}_2])$	$([\dot{s}_4,\dot{s}_4],[\ddot{s}_3,\ddot{s}_4])$	$([\dot{s}_4,\dot{s}_5],[\ddot{s}_1,\ddot{s}_1])$
a2	$([\dot{s}_4,\dot{s}_4],[\ddot{s}_2,\ddot{s}_3])$	$([\dot{s}_4,\dot{s}_5],[\ddot{s}_2,\ddot{s}_2])$	$([\dot{s}_1,\dot{s}_2],[\ddot{s}_3,\ddot{s}_4])$	$([\dot{s}_3,\dot{s}_3],[\ddot{s}_1,\ddot{s}_1])$
a3	$([\dot{s}_3,\dot{s}_4],[\ddot{s}_2,\ddot{s}_3])$	$([\dot{s}_5,\dot{s}_5],[\ddot{s}_2,\ddot{s}_2])$	$([\dot{s}_2,\dot{s}_3],[\ddot{s}_3,\ddot{s}_4])$	$([\dot{s}_4,\dot{s}_4],[\ddot{s}_1,\ddot{s}_1])$
a4	$([\dot{s}_3,\dot{s}_4],[\ddot{s}_2,\ddot{s}_3])$	$([\dot{s}_2,\dot{s}_3],[\ddot{s}_2,\ddot{s}_2])$	$([\dot{s}_4,\dot{s}_5],[\ddot{s}_3,\ddot{s}_4])$	$([\dot{s}_3,\dot{s}_5],[\ddot{s}_1,\ddot{s}_1])$

Table 4: aggregate attribute values by the 2DULWHMA and the rankings of the industry

p,q	industry	$\widetilde{d}_i(i=1,2,3,4)$	$E(\widetilde{d})$	Ranking
p=0.5,q=0.5	a1	$([\dot{s}_{0.188},\dot{s}_{0.336}],[\ddot{s}_1,\ddot{s}_1])$	0.01091	$a_1>a_2>a_4>a_3$
	a2	$([\dot{s}_{0.131},\dot{s}_{0.240}],[\ddot{s}_1,\ddot{s}_1])$	0.007745	
	a3	$([\dot{s}_{0.091},\dot{s}_{0.229}],[\ddot{s}_1,\ddot{s}_1])$	0.006672	
	a4	$([\dot{s}_{0.115},\dot{s}_{0.248}],[\ddot{s}_1,\ddot{s}_1])$	0.007554	
p=1,q=1	a1	$([\dot{s}_{0.324},\dot{s}_{0.375}],[\ddot{s}_1,\ddot{s}_1])$	0.014547	$a_1>a_2>a_4>a_3$
	a2	$([\dot{s}_{0.300},\dot{s}_{0.349}],[\ddot{s}_1,\ddot{s}_1])$	0.013526	
	a3	$([\dot{s}_{0.267},\dot{s}_{0.342}],[\ddot{s}_1,\ddot{s}_1])$	0.012687	
	a4	$([\dot{s}_{0.282},\dot{s}_{0.347}],[\ddot{s}_1,\ddot{s}_1])$	0.013122	
p=5,q=5	a1	$([\dot{s}_{0.765},\dot{s}_{0.770}],[\ddot{s}_1,\ddot{s}_1])$	0.0319827	$a_1>a_2>a_4>a_3$
	a2	$([\dot{s}_{0.763},\dot{s}_{0.768}],[\ddot{s}_1,\ddot{s}_1])$	0.031904	
	a3	$([\dot{s}_{0.759},\dot{s}_{0.767}],[\ddot{s}_1,\ddot{s}_1])$	0.031798	
	a4	$([\dot{s}_{0.761},\dot{s}_{0.768}],[\ddot{s}_1,\ddot{s}_1])$	0.031843	
p=10,q=10	a1	$([\dot{s}_{0.873},\dot{s}_{0.874}],[\ddot{s}_1,\ddot{s}_1])$	0.036389	$a_1>a_2>a_4>a_3$
	a2	$([\dot{s}_{0.872},\dot{s}_{0.873}],[\ddot{s}_1,\ddot{s}_1])$	0.036367	
	a3	$([\dot{s}_{0.871},\dot{s}_{0.873}],[\ddot{s}_1,\ddot{s}_1])$	0.036336	
	a4	$([\dot{s}_{0.872},\dot{s}_{0.873}],[\ddot{s}_1,\ddot{s}_1])$	0.036349	
p=1,q=0.5	a1	$([\dot{s}_{0.252},\dot{s}_{0.326}],[\ddot{s}_1,\ddot{s}_1])$	0.012046	$a_1>a_2>a_4>a_3$
	a2	$([\dot{s}_{0.218},\dot{s}_{0.286}],[\ddot{s}_1,\ddot{s}_1])$	0.010498	
	a3	$([\dot{s}_{0.180},\dot{s}_{0.277}],[\ddot{s}_1,\ddot{s}_1])$	0.009517	
	a4	$([\dot{s}_{0.199},\dot{s}_{0.285}],[\ddot{s}_1,\ddot{s}_1])$	0.010091	
p=1,q=2	a1	$([\dot{s}_{0.444},\dot{s}_{0.474}],[\ddot{s}_1,\ddot{s}_1])$	0.019121	$a_1>a_2>a_4>a_3$
	a2	$([\dot{s}_{0.430},\dot{s}_{0.460}],[\ddot{s}_1,\ddot{s}_1])$	0.018551	
	a3	$([\dot{s}_{0.407},\dot{s}_{0.455}],[\ddot{s}_1,\ddot{s}_1])$	0.017963	
	a4	$([\dot{s}_{0.417},\dot{s}_{0.458}],[\ddot{s}_1,\ddot{s}_1])$	0.01824	
p=1,q=5	a1	$([\dot{s}_{0.647},\dot{s}_{0.658}],[\ddot{s}_1,\ddot{s}_1])$	0.0272	$a_1>a_2>a_4>a_3$
	a2	$([\dot{s}_{0.643},\dot{s}_{0.654}],[\ddot{s}_1,\ddot{s}_1])$	0.02701	
	a3	$([\dot{s}_{0.633},\dot{s}_{0.652}],[\ddot{s}_1,\ddot{s}_1])$	0.02677	
	a4	$([\dot{s}_{0.637},\dot{s}_{0.653}],[\ddot{s}_1,\ddot{s}_1])$	0.026875	
p=1,q=10	a1	$([\dot{s}_{0.783},\dot{s}_{0.787}],[\ddot{s}_1,\ddot{s}_1])$	0.032726	$a_1>a_2>a_4>a_3$
	a2	$([\dot{s}_{0.782},\dot{s}_{0.786}],[\ddot{s}_1,\ddot{s}_1])$	0.03266	
	a3	$([\dot{s}_{0.778},\dot{s}_{0.785}],[\ddot{s}_1,\ddot{s}_1])$	0.03257	
	a4	$([\dot{s}_{0.780},\dot{s}_{0.785}],[\ddot{s}_1,\ddot{s}_1])$	0.032608	
p=0.5,q=1	a1	$([\dot{s}_{0.252},\dot{s}_{0.786}],[\ddot{s}_1,\ddot{s}_1])$	0.012046	$a_1>a_2>a_4>a_3$
	a2	$([\dot{s}_{0.218},\dot{s}_{0.286}],[\ddot{s}_1,\ddot{s}_1])$	0.010498	
	a3	$([\dot{s}_{0.180},\dot{s}_{0.277}],[\ddot{s}_1,\ddot{s}_1])$	0.009517	
	a4	$([\dot{s}_{0.199},\dot{s}_{0.285}],[\ddot{s}_1,\ddot{s}_1])$	0.010091	
p=2,q=1	a1	$([\dot{s}_{1.233},\dot{s}_{1.234}],[\ddot{s}_1,\ddot{s}_1])$	0.051394	$a_1>a_2>a_4>a_3$
	a2	$([\dot{s}_{1.106},\dot{s}_{1.160}],[\ddot{s}_1,\ddot{s}_1])$	0.047219	
	a3	$([\dot{s}_{0.917},\dot{s}_{1.106}],[\ddot{s}_1,\ddot{s}_1])$	0.042142	
	a4	$([\dot{s}_{1.039},\dot{s}_{1.164}],[\ddot{s}_1,\ddot{s}_1])$	0.045891	
p=5,q=1	a1	$([\dot{s}_{1.108},\dot{s}_{1.108}],[\ddot{s}_1,\ddot{s}_1])$	0.046167	$a_1>a_2>a_4>a_3$
	a2	$([\dot{s}_{1.050},\dot{s}_{1.074}],[\ddot{s}_1,\ddot{s}_1])$	0.044247	
	a3	$([\dot{s}_{0.958},\dot{s}_{1.050}],[\ddot{s}_1,\ddot{s}_1])$	0.041829	
	a4	$([\dot{s}_{1.016},\dot{s}_{1.075}],[\ddot{s}_1,\ddot{s}_1])$	0.043543	
p=10,q=1	a1	$([\dot{s}_{1.057},\dot{s}_{1.057}],[\ddot{s}_1,\ddot{s}_1])$	0.044041	$a_1>a_2>a_4>a_3$
	a2	$([\dot{s}_{1.027},\dot{s}_{1.039}],[\ddot{s}_1,\ddot{s}_1])$	0.043031	
	a3	$([\dot{s}_{0.977},\dot{s}_{1.027}],[\ddot{s}_1,\ddot{s}_1])$	0.041738	
	a4	$([\dot{s}_{1.008},\dot{s}_{1.039}],[\ddot{s}_1,\ddot{s}_1])$	0.04264	

Acknowledgment

This paper is supported by Social Sciences Planning Project of Shandong Province of China (13CGLZ02).

Bibliography

[1] Churchman C.W., Ackoff R.L., Arnoff E.L. (1975);*Introduction to operations research*, Wiley, New York, 1975.

[2] Wang X.Z., Dong C.R. (2009); Improving generalization of fuzzy if then rules by maximizing fuzzy entropy, *IEEE Transactions on Fuzzy Systems*, 17(3):556-567.

[3] Xi-Zhao Wang, Ling-Cai Dong, Jian-Hui Yan(2012); Maximum Ambiguity-Based Sample Selection in Fuzzy Decision Tree Induction. *IEEE Transactions on Knowledge & Data Engineering*,24(8):1491-1505.DOI:10.1109/TKDE.2011.67.

[4] Wang X.Z. et al. (2014); A study on relationship between generalization abilities and fuzziness of base classifiers in ensemble learning, *IEEE Transactions on Fuzzy Systems*, 23(5):1638 - 1654. DOI:10.1109/TFUZZ.2014.237147.

[5] Bingsheng Liu et al. (2014); A complex multi-attribute large-group decision making method based on the interval-valued intuitionistic fuzzy principal component analysis model.*Soft Comput*, 18:2149-2160, DOI 10.1007/s00500-013-1190-8.

[6] M. Delgado, J.L. Verdegay, M.A. Vila (1993); On aggregation operations of linguistic labels, *International Journal of Intelligent Systems*,8:351-370.

[7] F. Herrera-Martinez(2000); A 2-tuple fuzzy linguistic representation model for computing with words, *IEEE Transactions on Fuzzy Systems*,8(6):746-752.

[8] Liu P.D., Chen Y.B., Chu Y.C., (2014); Intuitionistic uncertain linguistic weighted bonferroni owa operator and its application to multiple attribute decision making, *Cybernetics and Systems*, 45(5):418-438, DOI:10.1080/01969722.2014.929348.

[9] Liu P.D., Chu Y.C., Li Y.W. (2015); The multi-attribute group decision-making method based on the interval grey uncertain linguistic variable generalized hybrid averaging operator, *Neural Comput Appl*, 26(6):1395-1405.

[10] Liu X. et al. (2016); A New Interval-valued 2-Tuple Linguistic Bonferroni Mean Operator and Its Application to Multiattribute Group Decision Making, *Int. J. Fuzzy Syst*,12: 1-23.DOI:10.1007/s40815-015-0130-4.

[11] Liu P.D., Chen Y.B., Chu Y.C. (2014); Intuitionistic uncertain linguistic weighted bonferroni owa operator and its application to multiple attribute decision making, *Cybernetics and Systems*, 45(5):418-438, DOI:10.1080/01969722.2014.929348

[12] Liu P.D., Shi L.L. (2015); Intuitionistic uncertain linguistic powered einstein aggregation operators and their application to multi-attribute group decision making, *J Appl Anal Comput*, 5(4):534-561.

[13] Liu P.D., Wang Y.M. (2014); Multiple attribute group decision making methods based on intuitionistic linguistic power generalized aggregation operators. *Appl Soft Comput*, 17:90-104.

[14] Liu P.D., Yu X.C. (2014); 2-Dimension uncertain linguistic power generalized weighted aggregation operator and its application in multiple attribute group decision making, *Knowl Based Syst*, 57:69-80.

[15] Zhu W.D., Zhou G.Z., Yang S.L. (2009); An approach to group decision making based on 2-dimension linguistic assessment information, *Syst Eng*, 27:113-118.

[16] Liu P.D., Zhang X. (2012); An Approach to Group Decision Making Based on 2-dimension Uncertain Linguistic Assessment Information, *Technol Econ Dev Econ*, 18(3):424-437.

[17] Liu P., He L., Yu X. (2016); Generalized Hybrid Aggregation Operators Based on the 2-Dimension Uncertain Linguistic Information for Multiple Attribute Group Decision Making, *Group Decision and Negotiation*,25(1):103-126, DOI:10.1007/s10726-015-9434-x

[18] Liu P.D., Qi X.F. (2014); Some generalized dependent aggregation operators with 2-dimension linguistic information and their application to group decision making, *J Intell Fuzzy Syst*, 27:1761-1773.

[19] Liu P., Teng F.(2016); Multiple attribute decision-making method based on 2-dimension uncertain linguistic density generalized hybrid weighted averaging operator, *Soft Comput*, 10(3): 1-14, DOI 10.1007/s00500-016-2384-7.

[20] Liu P.,Wang Y.(2015); The aggregation operators based on the 2-dimension uncertain linguistic information and their application to decision making.*Mach. Learn. & Cyber*, 1-18, DOI:10.1007/s13042-015-0430-x.

[21] Yu X.H., Xu Z.S., Liu S.S., Chen Q. (2012); Multi-criteria decision making with 2-dimension linguistic aggregation techniques, *Int J Intell Syst*, 27:539-562.

[22] P.D. Liu, X. Zhang(2012); Intuitionistic uncertain linguistic aggregation operators and their application to group decision making, *Systems Engineering-Theory & Practice*, 32(12): 2704-2711.

[23] Z.S. Xu (2006); Goal programming models for multiple attribute decision making under linguistic setting, *Journal of Management Sciences in China*, 9(2): 9- 17.

[24] Xu Z. (2006); A note on linguistic hybrid arithmetic averaging operator in multiple attribute group decision making with linguistic information. *Group Decision and Negotiation*, 15(6), 593-604.

[25] Beliakov G., Pradera A., Calvo T. (2007); Aggregation Functions: A Guide for Practitioners, *Springer Berlin Heidelberg.*

[26] Yu D. (2013); Intuitionistic fuzzy geometric heronian mean aggregation operators, *Applied Soft Computing*, 13(2): 1235-1246, DOI: 10.1016/j.asoc.2012.09.021

INTERNATIONAL JOURNAL OF COMPUTERS COMMUNICATIONS & CONTROL
ISSN 1841-9836, 12(2):265-275, April 2017.

Balancing Between Exploration and Exploitation in ACO

A.E. Negulescu, S.C. Negulescu, I. Dzitac

Alina Eugenia Negulescu
University POLITEHNICA of Bucharest
Romania, 060042 Bucharest, sector 6, Splaiul Independentei, 313
alina.lascu@gmail.com

Sorin Constantin Negulescu
Lucian Blaga University of Sibiu
Romania, 550024 Sibiu, Victoriei Bvd., 10
sorin.negulescu@ulbsibiu.ro

Ioan Dzitac
1. Aurel Vlaicu University of Arad
Romania, 310330 Arad, Elena Dragoi, 2
ioan.dzitac@uav.ro
2. Agora University of Oradea
Romania, 410526 Oradea, Piata Tineretului, 8
rector@univagora.ro

Abstract: In order to balance the preference of the artificial entities towards *exploration* or *exploitation* (in their transition rule), a novel technique is proposed for replacing the random function used by the classical Ant Colony Optimization (ACO) algorithms for solving the Traveling Salesman Problem (TSP). The proposed Beta Distribution function (**B**), or $random.betavariate(a, b)$ has the proven capability (depicted through test-runs) of influencing the algorithm's solution quality and convergence speed. Consequently, this paper will introduce in the related work section the classical ACO algorithm, with a focus on the *transition rule* used for choosing the next node in the problem's associated graph, followed by the related work on this topic, and it will continue with the introduction of the **B** function which will be presented both from a theoretical and practical perspective in relation with the scope: balancing between exploration and exploitation in order to improve the performance of the ACO algorithm for the TSP. The paper concludes that the B-EAS has the ability to find better solution than EAS for a set of benchmarks from the TSPLib library.
Keywords: ACO, TSP, **B**, Elitist Ant System (EAS), Beta-Elitist Ant System (B-EAS), transition rule.

1 Introduction

Swarm inspired algorithms, and specifically the ACO area, represent a steady interest over the past five years, as reflected in searches and recorded by Google Trends [7]. The constant focus on this family of algorithms is due to their simplicity in terms of modeling and, implicitly, the ease of understanding their functionality, which makes them, in turn, very appealing for computer scientists aiming to solve optimization problems.

Since they are inspired from nature, the principles of the ACO algorithms are easy to grasp. Their well-known applicability is for solving the TSP, which can also be very easily explained but, unfortunately, not that simple to solve: a traveling salesman has a list of cities which he must visit and return back to the starting point; there are several conditions that must be met: he must pass through each city only once, and he must find the shortest route to fulfill this task, as

Copyright © 2006-2017 by CCC Publications

otherwise, longer routes will be associated with bigger costs. This problem can be extrapolated to the world of ants which, foraging for food, need to find the shortest path between the nest and the food source. Because of these resemblances in terms of tasks, the ant colony inspired algorithms are directly applicable for solving the TSP.

Pertaining the parameters, by mirroring the natural ant colony model, the classical ACO algorithm operates with artificial *entities* (i.e. the ants), *positive feedback* using pheromone deposit in the environment, *heuristic information* regarding distances in the search-space and *negative feedback* resulted from pheromone evaporation. For all intents and purposes, these parameters, which are inspired from nature, are the most influential ones in any ACO algorithm.

Other common traits between ACO algorithm variants are the initialization steps that include:

- loading the problem's associated graph;
- setting up the initial parameter's values;
- initialization of the pheromone trails on the graph's arches (using a minimum quantity of pheromone) as to "kick-start" the algorithm as otherwise, it would take a couple of iterations until the artificial ants would do this instead and this would result in additional computational costs.

Even though the last described step is obviously not characteristic to real ants, the model has been slightly altered so as to increase the convergence speed of the algorithm towards a solution.

After performing the initialization, as long as the algorithm's stop condition is not met (represented by, for instance, a maximum number of iterations or time limitation), every artificial ant needs to perform a set of steps, part of the construct solutions function:

- *choose a starting node on the graph* – predictably (ant 1 in node 1, ant 2 in node 2 and so on) or randomly;
- *choose next node* until all nodes are visited only once (function on which this paper is centered), during which the "artificial ant has to decide" which nodes are more attractive and choose one of them based on a probability formula that will be presented shortly;

After all artificial ants have constructed solutions:

- *update pheromone trail* function is employed for depositing a pheromone quantity on all the arches of the graph visited by the artificial ants during a tour, which, following the natural model, represents the positive feedback, or reinforcement; in this way the arches that have been used by multiple ants will have a greater pheromone intensity;
- *evaporate pheromone trail* function will act as negative feedback in order for the algorithm to "forget" old and inefficient trails; this step is paramount as without it, the algorithm would converge too fast and get trapped in local optima;
- *store best solution* function will compare all tour lengths of the artificial ants and record the smallest length value along with the order in which the nodes have been visited, for future comparisons.

Having presented the context, the paper will focus on the transition rule of ACO algorithms which is based on a probability formula (as mentioned earlier). The state of the art is presented in "Origins and related work", while the "Rationale and approach" section pinpoints the drawbacks of current approaches and presents the *beta distribution function* as a possible solution. The

section entitled "The model", describes the changes that have been applied to the EAS algorithm in order to support the identified approach for improvement: Beta-Elitist Ant System (B-EAS) algorithm. The "Conclusions and intentions" section rounds up the paper by presenting the strengths and weaknesses of the B-EAS algorithm, together with the identified future research intentions in this direction.

2 Origins and related work

Initial experiments, conducted with real ants (*Linepithaema humile*), proved that when ants have to choose between two possible paths of different lengths (when going between a food source and the nest) they will initially travel on any of them with an equal probability but, as time passes, the shorter one will have a higher amount of deposited pheromone (reinforced by multiple, shorter trips) whereas the longer one will have a lower amount of pheromone (due to fewer, longer trips). Moreover, considering that the evaporation speed of the pheromone is equal, the intensity of the pheromone on the longer path will decrease faster. Even though the pheromone intensity on the two paths was different, a very small number of ants were still choosing the one with smaller pheromone intensity from time to time, which lead to the conclusion that the ants were deciding in a *probabilistic* manner.

The findings in this "double-bridge experiment" [8] were adopted when the initial Ant System (AS) algorithm was modeled [5]. As such, almost all ACO algorithms are using a probability formula for determining the next node. The formula coined by AS, which is reproduced below (unaltered) defines, as called by its authors, a *random-proportional rule*:

$$p_{i,j}^{k}(t) = \begin{cases} \frac{[\tau_{ij}(t)]^{\alpha} \cdot [\eta_{ij}(t)]^{\beta}}{\sum_{l \in N_i^k} [\tau_{il}(t)]^{\alpha} \cdot [\eta_{il}(t)]^{\beta}} & if j \in N_i^k \\ 0 & if j \notin N_i^k \end{cases} \quad (1)$$

Formula 1 corresponds to the probability p at the moment t for an ant k visiting node i to choose node j as the next node in which it will go. α and β are two general parameters used for weighting the pheromone intensity τ_{ij} and the visibility η_{ij} (defined as $1/distance$) between nodes i and j when computing (at the moment t) the cost of choosing node j. The set N_i^k contains the nodes the ant k is allowed to choose (i.e. all the nodes which were not yet visited, as constrained by the TSP). So, the formula can be expressed as being the ratio between the cost of choosing node j and the sum of costs for all the other nodes not yet visited. Of course, this probability is 0 if node j is already visited.

Before the above presented formula is applied, a random real number between 0 and 1 is generated, representing the probability value. Afterwards, $p_{ij}^{k}(t)$ is calculated for as many nodes $j \in N_i^k$ as necessary, until the sum of these computed costs is equal or greater than the selected probability and the last j node is chosen to move to. In this way, in many cases the formula is not applied for all the nodes not yet visited, sparing some time. The steps above are repeated until no more nodes are left to visit, resulting a list of nodes visited by the ant k.

As a further improvement of the initial AS algorithm, Dorigo and Gambardella proposed in the Ant Colony System (ACS) algorithm a new transition rule that included *exploration* and *exploitation* traits [6]. This improvement, as described by its authors, is due to the *pseudo-random-proportional rule* for choosing the next node in the itinerary, where the ant k in the node i selects the next node j with a probability of $0\% \leq q_0 \leq 100\%$. As such, a q_0 value is a priori selected when the algorithm starts, whereas the next node is selected by applying one of the following formulas depending on a randomly generated value q:

- if $q < q_0$ then the node for which $\tau_{ij}(t) \cdot [\eta_{ij}(t)]^{\beta}$ is maximum, is used;

- else, the node is selected by using the same formula introduced in the AS algorithm.

As described in [6], "ACS algorithm's state transition rule provides a direct way to balance between *exploration of new edges* and *exploitation of a priori and accumulated knowledge* about the problem". Along with other several improvements, ACS' proven efficiency was demonstrated by its authors through a series of experiments whose results had shown that this algorithm outperformed other algorithms inspired from nature, as for instance, simulated annealing and evolutionary computation.

Another improvement of the AS was the Elitist Ant System (EAS) [5] that was reinforcing the pheromone intensity on the arches belonging to the tour of the ants that found the best solutions at the moment t. In this way one can state that the elitist ants can deposit a greater amount of pheromone. The experiments of the before mentioned authors indicated that a relatively small number of elitist ants can help the performance of the algorithm.

Nevertheless, as depicted in [1], the author claims that "the transition rule suggested in the original Ant System and used extensively in modified versions [...] uses a power law scaling procedure to define the relative importance of the pheromone intensity and local information. [...] This form of transition rule, however, often leads to premature convergence of the ant algorithms to suboptimal solution when used with elitist strategy of pheromone updating". The author proposed, as a result, an alternative transition rule to be used with elitist strategy of pheromone updating in ant algorithms. Moreover, the author of [1] considers that the classical transition rule presented in formula 1 would be the source of the local optima issue which called for the design of the MAX-MIN Ant System (MMAS) algorithm. Further explained in [1], the reason for which the transition rule is not efficient is that the influence of the pheromone intensity (α) will play a main role in the decision making process of an ant choosing the next node to visit, while the heuristic information will have no impact. This is determining, as a result, the stagnation of the search. Consequently, the author proposes an alternative method of transition that would prevent the pheromone intensity from "dominating the ants decision table at all stages of computation". This formula is represented as:

$$p_{i,j}^{k}(t) = \begin{cases} \frac{\alpha\tau_{ij}(t)\cdot\beta\eta_{ij}}{\sum_{l\in N_i^k[\alpha\tau_{il}(t)\cdot\beta\eta_{il}]}} & if j \in N_i^k \\ 0 & if j \notin N_i^k \end{cases} \tag{2}$$

The author in [1] claims that, this new transition rule based on addition, will ensure that for "the properly adjusted values of the pheromone, and local heuristics sensitivity parameters, the local heuristic term η_{ij} will always have a saying in forming the decision table, irrespective of the distribution of pheromone intensities". Moreover, this method has the advantage of "reducing the number of controlling parameters, as one can always assume the value of unity for one of these parameters and adjust the value of the other one for the best performance of the method", as described by the aforementioned author.

Other solutions that were targeting the improvement of the original AS algorithm were the Rank-Based Ant System (RBAS) [3], Best-Worst Ant System (BWAS) [4], HyperCube-Ant Colony Optimization (HC-ACO) [2] and Grahp-Based Ant System (GBAS) [9], to name a few of the most common. Also, a respectable number of approaches went beyond the initial paradigm, implementing pheromone correction strategies [14] or creating hybrid solutions involving evolutionary computation [15] and neural networks [13].

Whereas in the classical ACO algorithm, α and β parameters are global, in the model presented in [11], they become local and "personal" to each individual ant. In this way, ants can behave differently in their transition rule. Considering that the resource costs associated with this approach is minimal (in terms of memory) the benefits are various: this variant of ACO

algorithm has the ability to self-tune depending on the problem type and size (if local α and β parameters are adjusted based on the tour lengths found by each ant), the convergence speed can be improved and the parameters values can be easily adjusted automatically. All these benefits can assist in addressing the stagnation issue (i.e. local optima) that is typical to all meta-heuristic algorithms.

3 Rationale and approach

As it can be depicted from the selected examples provided in the previous section, there is an on-going interest regarding the improvement of the ACO algorithms, due to their proven potential for solving combinatorial optimization problems. Some improvement approaches are focusing on the *update pheromone intensities rule*, others are looking towards the parameters' values setting, while this paper investigates the *transition rule* which the ants use to move through the nodes in the graph.

All algorithms from the meta-heuristic class are relaying on a probability that implies the use of randomly generated numbers in different manners. Without this intrinsic randomness factor the behavior would tend to be the same as the best-first algorithm and the solutions quality would be poor.

The approach presented in this paper, that would allow to change the algorithm's behavior towards exploration or exploitation, is based on another type of probability: **B**, that would allow a *beta-proportional rule*.

Beta distribution can be defined, according to [10] as "a mean of distribution probabilities", that essentially, represents "all the possible values of an unknown probability". This method is reasonably employed by statistics and probability theory and is defined in $[0, 1]$ domain, having two positive shape parameters: a and b. Its application is widely used in diverse areas as a technique that allows influencing the behavior of random variables limited to ranges of limited lengths. Given these features, it can be used as a proper model for random behavior (pertaining the uncertainty of an outcome).

For all practical purposes, a test-run may only have two possible results: either success (with a probability of x) or failure (with a probability of $1 - x$). As explained by the same author, this can be probabilistically illustrated by assigning to x a uniform distribution in the $[0, 1]$ domain, as x being a probability, it can only take values between 0 and 1. Also, the uniform distribution assigns respective probability density to every entity in the domain, which effectively means, that no possible value of x is *a priori*, more likely than the other. As such, for n independent iterations performed during an experiment, there can result both k successful outcomes, as well as $n - k$ failed ones. Consequently, the conditional distribution of x, based on the observation of k successful outcomes resulted from n iterations, represents, actually, a beta distribution with $k + 1$ and $n - k + 1$ parameters.

Therefore, as depicted in [10], for x, absolutely continuous random variable: $R_x = [0, 1]$ and $a, b \in \mathbb{R}_{++}$, x has a beta distribution with a, b as shape parameters of x's probability density function, presented through the following formula, where $\mathbf{B}(a, b)$ represents the **B** function:

$$f(x) = \begin{cases} \frac{1}{B(a,b)} x^{a-1} \left(1 - x^{b-1}\right) & if x \in R_x \\ 0 & if x \notin R_x \end{cases} \tag{3}$$

To visually grasp the behavior of this function for different parameter values, below are presented several histograms. The data set is represented by 10000 samples with values within the $[0, 500]$ domain on the abscisa and the number of randomly generated samples is displayed on the ordinate.

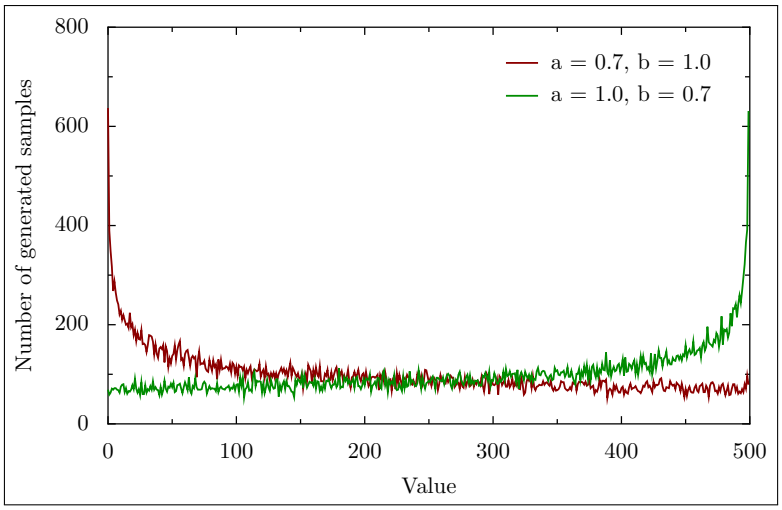


Figure 1: Histogram for beta distribution function with one sub-unitary and one unitary argument

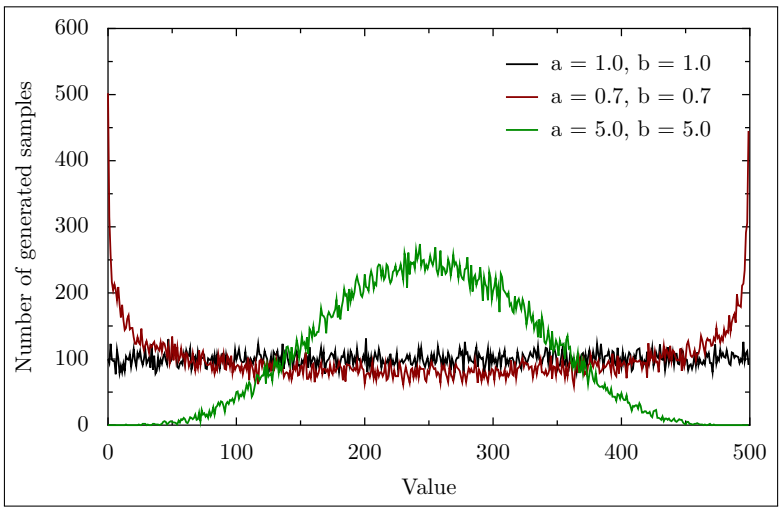


Figure 2: Histogram for beta distribution function with equal arguments values

According to the visual representation of **B** in figures 1 and 2, one can easily depict the fact that this function is highly flexible since it allows, depending on the values of its parameters, to randomly generate a higher number of values at one end or the other of the interval, a uniform distribution and a normal (inverse) distribution. Also, it is important to mention that, if the a and b parameters equal 1, this function is behaving exactly like the classic random function where all the values have an equal probability of being generated and the algorithm would perform exactly like the AS algorithm.

4 The model

Based on the previously introduced approach, consisting in replacing the *random-proportional* transition rule with the *beta-proportional* one, this section describes in detail, the model and the method used for testing it. The foundation of the model is the EAS algorithm where the herein proposed transition rule has been modified to include the following steps:

- Determine the costs of moving from the current node to all the nodes not yet visited by applying the formula for computing $p_{ji}^{k}(t)$ using the same formula used by EAS (as presented in the origins and related work section).
- Store these costs together with the corresponding node number in a vector by inserting them in such a way that the vector is ordered increasingly (so the nodes that have smaller costs are at the beginning of the vector). Obviously, generating this vector in an ordered fashion has the advantage that it is much faster than generating and then sorting it.
- Generate a random number according to **B** that will indicate the position (in the vector) of the node selected as the next node to move to. The values of the parameters for the function must be chosen in such a way that it will generate with a higher probability small values so the elements at the beginning of the vector have a higher chance to be selected.

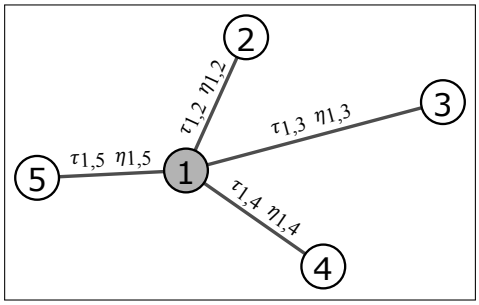


Figure 3: Example of associated graph with heuristic and pheromone information

For a better understanding of the steps involved in the new model called Beta-Elitist Ant System (B-EAS), the graph in figure 3 is provided as an example. Presuming that an ant starts in node 1, the cost of going to nodes $2, 3, 4$ and 5 is taking into consideration the pheromone intensity and the visibility on each arch.

The ordered vector, resulted after computing the costs, is presented in figure 4, together with the probability of selection according to **B** function. As it can be depicted in the example, node 3 which involves the smallest cost in terms of visibility and pheromone intensity, has a higher probability of being selected than $2, 4$ and 5 if the arguments of **B** are properly chosen.

Test-runs were conducted for a full factorial experiment in order to determine appropriate values for the a and b parameters of **B**. Subsequently, in order to generate smaller values with

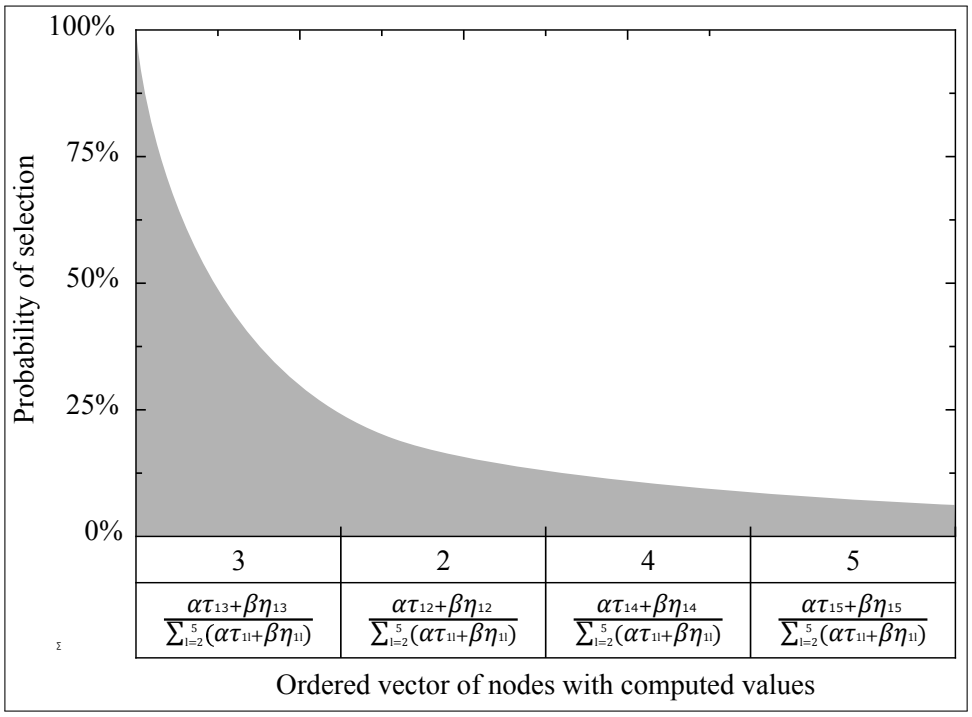


Figure 4: Example of selection probability for the ordered vector's elements

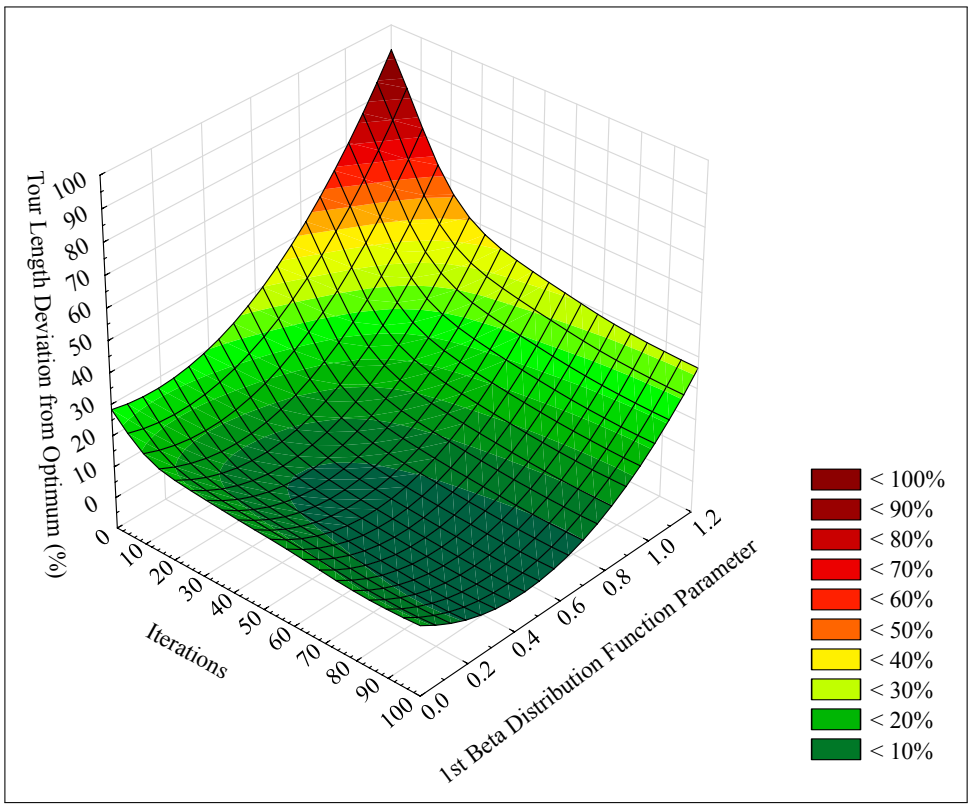


Figure 5: Solution length versus 1^{st} parameter of beta distribution function)

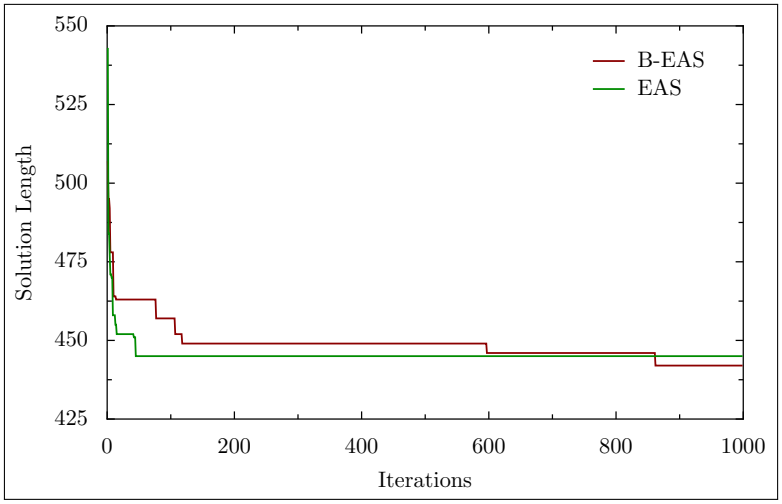


Figure 6: Comparison between EAS and B-EAS algorithms running on eil51 benchmark

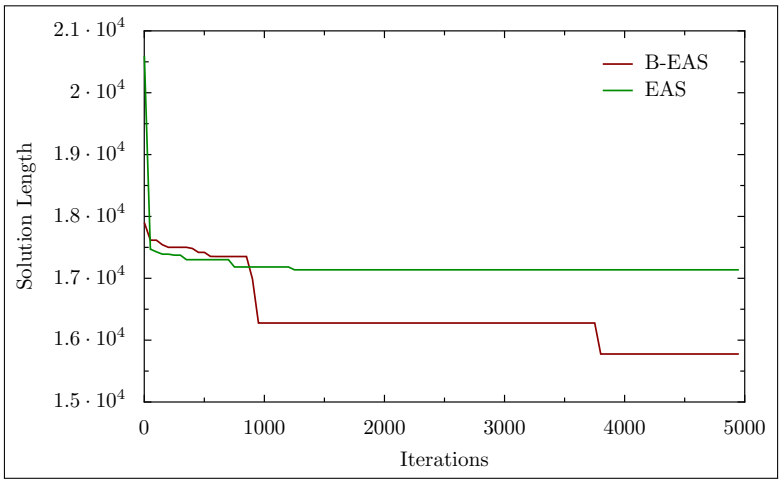


Figure 7: Comparison between EAS and B-EAS algorithms running on d198 benchmark

a higher probability, the values for parameter a have been tested within $[0.1, 1]$ domain, while b was maintained constant at 1. The results are presented further on in figure 5, where, for a set of problems with known solutions, the algorithm was constrained to run for 100 iterations. The quality of the obtained solutions were averaged and compared in a percentual manner to the known solutions so, the optimal solution obtained is 0% deviating from the best known ones (represented in darker color in the graph).

The conclusion of this experiment is that the optimal parameters' values of **B** function, used for generating random numbers, are $a = 0.4$ and $b = 1$. As displayed in figure 5 (where the dependent variable *tour length* is expressed versus parameter a and the *number of iterations* dependent variables), these values are driving the algorithm towards finding better solutions (with shorter lengths), and with a higher convergence speed.

A comparison between the EAS and B-EAS algorithms using standard benchmarks from the TSPLib [12] in regards to solutions quality and convergence speed is presented in figures 6 and 7. We have observed that without modifying the a and b parameters during the test-runs, the B-EAS was converging slower than EAS, but it continued to find better solutions, whereas EAS would have stopped in a local optima.

As stated before, when parameters have the values $a < 1$ and $b = 1$, ants tend to choose with a higher probability the nodes that give smaller costs of travel (*exploitation*), but there is also the probability to choose nodes that imply higher costs (*exploration*). This is why the graphs are displaying large improvements that are not further more developed over a high number of iterations, especially when a is much smaller than 1.

Conclusions and intentions

The paper presents a new algorithm (B-EAS) that can be stirred towards *exploitation* or *exploration* by modifying the a and b parameters used in the *beta-proportional transition rule*. Specifically, the hereby authors propose replacing the classical random function for choosing the next node with the beta distribution function which allows a more flexible way of generating random numbers.

To validate the proposed model, the optimal parameter values of **B** were determined for a set of benchmark problems using a full factorial experiment. The next logical step was to compare through test-runs the performance of B-EAS with the original algorithm from which it was derived (EAS). The results have shown that B-EAS finds better solutions than EAS, given enough time or iterations.

Considering the approach introduced by the same authors in [11], where the global parameters of the algorithm (α, β and ϵ) were transformed into local parameters for each artificial ant, creating thus a "population" of agents with individual characteristics (personalities), the same approach could be applied for the a, and b parameters of the **B** function, as well. We consider that this research direction has the potential to create a genetic-featured ACO, where the algorithm could evolve and "tune" itself.

Bibliography

[1] Afshar, M.H. (2005); A new transition rule for ant colony optimization algorithms: application to pipe network optimization problems, *Engineering Optimization*, 37 (5): 525-540.

[2] Blum, C.; Roli, A.; Dorigo, M. (2001); HC-ACO: The Hyper-Cube Framework for Ant Colony Optimization, *IEEE Transactions on Systems, Man, and Cybernetics-Part B*, 399-403.

[3] Bullnheimer, B.; Hartl, R.F.; Strauss, C. (1999); A New Rank Based Version of the Ant System: A Computational Study, *Central European Journal for Operations Research and Economics*, 25-38.

[4] Cordón, O.; de Viana, I.F.; Herrera, F.; Moreno, L. (2000); A new ACO model integrating evolutionary computation concepts: The best-worst Ant System. *In M. Dorigo, M. Middendorf, & T. Stuützle (Eds.), Abstract Proceedings of ANTS 2000-From Ant Colonies to Artificial Ants: Second International Workshop on Ant Algorithms*, Brussels, IRIDIA, Université Libre de Bruxelles, 22-29.

[5] Dorigo, M.; Maniezo, V.; Colorni, A. (1996); Ant system: optimization by a colony of cooperating agents, *IEEE Transactions on Systems, Man, and Cybernetics - Part B*, 26 (1): 29-41.

[6] Dorigo, M.; Gambardella, L. (1997); Ant Colony System: A Cooperative Learning. *IEEE Transactions of Evolutionary Computation*, 1 (1): 53-66.

[7] Google (2016). Google Trends, Retrieved from `https://www.google.com/trends/explore?q=AntColonyOptimization`.

[8] Goss, S.; Aron, S.; Deneubourg, J.L.; Pasteels, J.M. (1989); Self-organized shortcuts in the Argentine ant, *Naturwissenschaften*, 76, 579-581, Springer-Verlag.

[9] Gutjahr, W.J. (2000). A Graph-based Ant System and its convergence, *Future Generation Computer Systems*, 873-888.

[10] Kerns, G. (2011); Introduction to Probability and Statistics Using R, *IPSUR*, ISBN: 978-0-557-24979-4.

[11] Negulescu, S.; Dzitac, I.; Lascu, A. (2010); Synthetic Genes for Artificial Ants. Diversity in Ant Colony Optimization Algorithms, *International Journal of Computers Communication & Control*, 5 (2): 216-223.

[12] Reinelt, G. (1991); TSPLIB - A Traveling Salesman Problem Library. *ORSA Journal on Computing*, 376-384.

[13] Susnea, I.; Axenie, C. (2015); Cognitive maps for indirect coordination of intelligent agents, *Studies in Informatics and Control*, 24(1): 111-118.

[14] Tuba, M.; Jovanovic, R.(2013); Improved ACO Algorithm with Pheromone Correction Strategy for the Traveling Salesman Problem, *International Journal of Computers Communications & Control*, 8(3): 477-485.

[15] Wei, X.; Han, L.; Hong, L. (2014); A modified ant colony algorithm for traveling salesman problem, *International Journal of Computers Communications & Control*, 9(5): 633-643.

INTERNATIONAL JOURNAL OF COMPUTERS COMMUNICATIONS & CONTROL
ISSN 1841-9836, 12(2):276-290, April 2017.

A Solution for Problems in the Organization, Storage and Processing of Large Data Banks of Physiological Variables

F. Palominos, H. Díaz, F. Córdova, L. Cañete, C. Durán

Fredi Palominos*, Hernan Díaz, Lucio Cañete, Claudia Durán
Departaments: Mathematics, Biology, Industrial Technology and Industrial Engineering.
Universidad de Santiago de Chile
Avda. Bernardo O'Higgins 3363, Santiago, Chili.
(fredi.palominos, hernan.diaz, lucio.canete, claudia.duran)@usach.cl
*Corresponding author: fredi.palominos@usach.cl

Felisa Córdova
Director at Engineering School
Finis Terrae University
felisa.cordova@gmail.com

Abstract: The proliferation and popularization of new instruments for measuring different types of electrophysiological variables have generated the need to store huge volumes of information, corresponding to the records obtained by applying this instruments on experimental subjects. Together with this must be added the data derived from the analysis and purification processes. Moreover, several stages involved in the processing of data is associated with one or more specific methods related to the area of research and to the treatment at which the base information (RAW) is subjected. As a result of this and with the passage of time, various problems occur, which are the most obvious consequence of that data and metadata derived from the treatment processes and analysis and can end up accumulating and requiring more storage space than the base data. In addition, the enormous amount of information, as it increases over time, can lead to the loss of the link between the processed data, the methods of treatment used, and the analysis performed so that eventually all becomes simply a huge repository of biometric data, devoid of meaning and sense. This paper presents an approach founded on a data model that can adequately handle different types of chronologies of physiological and emotional information, ensuring confidentiality of information according to the experimental protocols and relevant ethical requirements, linking the information with the methods of treatment used and the technical and scientific documents derived from the analysis. Consequently, the need to generate specific data model is justified by the fact that the tools currently associated with the storage of large volumes of information are not able to take care of the semantic elements that make up the metadata and information relating to the analysis of base records of physiological information. This work is an extension of our paper [25].[a]
Keywords: data models; big data; EEG data organization; physiological information, metadata.

[a]Reprinted (partial) and extended, with permission based on License Number 3947080516854 [2016] ©IEEE, from "Computers Communications and Control (ICCCC), 2016 6th International Conference on".

1 Introduction

This paper is an extension of [25] which delves into the specification of the different types of relevant information on the EEG records, the mathematical formalization of data and interactions expands, and indicators to quantify and predict storage requirements are proposed.

Understanding the mechanisms of human reasoning and brain functioning is a central topic in neuroscience [16] [17] [18] [30] . Technological development has multiplied alternatives of

Copyright © 2006-2017 by CCC Publications

technological tools to record brain activity. In turn, the development of computers has created new opportunities for processing and analyzing the data through different tools [19] [20] [21]. Moreover, the greater availability and lower prices of technology allow researchers to access many instrumental that was previously reserved for medical specialists, allowing researchers also raised new objectives and scientific concerns.

As a result, in addition to the generation of new knowledge, there has been a huge growth in the quantity of EEG records, generating huge volumes of research and clinical data, in centers worldwide. The enormous size of these records, evidence that it is a collection of large volumes of data in the field of big data, which to be registered in computer systems, requires enormous storage media, which escape the most feature traditional computer systems.

The origin of the electroencephalogram, an instrument that allows the capture of signals known as EEG, date back to 1875 when Richard Caton first detected electrical activity on the brain surface of animals [22]. However, the expanding use of this technology beyond the traditional medical field occurs only in the last decades together with new quantitative approaches (qEEG) allowing a deeper data comprehension beyond qualitative characterizations. [23].

The potential to use this technology spans multiple areas, such as neuromarketing [24], education [26], psychology [27], labor relations [28] and work [29] .

1.1 The EEG Data and theirs specificities

Electroencephalographic data (EEG) are the record obtained by measuring brain activity, using the instrument called electroencephalographs. The measurements are obtained through a set of electrodes that are located at certain points of the human scalp, generally non-invasively, such that each electrode receives an analog signal which is then digitized by a computer. Because of the nature of the EEG data recording and digitization procedure, the temporal resolution of the signal will only be limited by the sample rate used in the experiment.

While a good amount of knowledge has been gathered from an ample range of EEG frequency span (mainly between 0,5 and 30 Hz), which mean between the EEG wave ranges of delta and beta, the gamma band (>30 Hz) has been scarcely explored.

To have a look at gamma frequencies it is necessary to increase the sample rate up to 256 Hz to have a confident resolution measuring brain phenomena that occur at 128 Hz maximum. If we want to go further, we only need to double the sample rate to 512 Hz to have a new maximum resolution of 256 Hz for the brain phenomena.

Starting with a standard EEG configuration of sample rate 128 Hz, we can explore the initial part of the gamma oscillation between 30 Hz (the end of beta) until 64 Hz (the maximum resolution allowed by Nyquist theorem). By increasing our sample rate to 256 Hz we increased data file from 64 data points in a second to 128 data points.

By setting the sample rate to 512 Hz, we ended with 256 data points per second. The fractal nature of the EEG signal makes it only finite when our technology to study it fails to go beyond.

To reach real time performances in brain-machine communication or to study very precise stimulus-response experiments, sample data must be recorded at thousands of Hz with the purpose of seeing deeper into the different time-dimensions of the processes happening in the brain.

The application of ECG in humans, although it is conditioned to the experimental protocols of each investigation, which has management, storage and processing of data concerned, have common characteristics. Sequences series of values recorded during application of an electroencephalogram (EEG), can be described as a set of analog sequences, called channels, such that each channel corresponds to record one of the electrodes of the electroencephalograph. The number of channels varies according to the instrument. The duration of each record depends on the experimental design and conditions. However, each channel will have the same duration as the

remaining channels of the RAW register.

The storage format may vary depending on the instrument type. In the case of the data corresponding to this job, they are stored in EDF format [1]. The length of each record should be as the experiment or clinical study established.

Due to the nature of brain activity, and the scope of the EEG technology, RAW records are not completely clean and can be affected by different perturbations. In the clinical field are called artifacts (residual noise from the instruments, interactions between channels, and twitching or muscle movements).

1.2 Problems on the EGG Data Management

Regarding the different factors related to storage, organization, processing and analysis of EEG records, it is clear that this is a complex problem of data management. The constant and often excessive amount of growth data and great accumulation of metadata generated in the processing and analysis of raw records. Loss of the meaning of information as a result of growth in the amount of data that over time hidden links between data, data processing, and results.

Many physiological data should be organized according to the time in which they originate, whether as instrumental records or as a result of processing. Therefore, proper management of the timestamps associated with the data is required.

In order not to incur losses of information and meaning, it is required to manage data in conjunction with information on related research, researchers, experimental people, derived data and results. Moreover, it is necessary to ensure the confidentiality of information.

Finally, it is necessary to record information about the nature of the data, their growth potential, the way they have been processed and the results of the analyses performed.

1.3 The EEG records and their relation to the area of Big Data

Current alternatives to storing and distributing data from the scope of Big Data are still weak in their ability to include metadata and meaning.

'Big Data refers to enormous amounts of unstructured data produced by high-performance applications falling in a wide and heterogeneous family of application scenarios' [7] . On the other hand, the problem of big data can be discerned as two distinct problems: Big data collections and Big Data objects [8] . 'There are three fundamental issue areas that need to be addressed in dealing with big data: storage issues, management issues, and processing issues' [12] and Big Data features can be summarized as follows: Data volume; Data velocity; Data variety; Data Value and Complexity [12] [10] .

The problems in the area of Big Data arise mainly from the needs of scalability, the massive scale of the data, the heterogeneity of information, unstructured data and their distribution across multiple platforms. Resulting in problems of management, storage, portability and processing of information. Because data can be distributed across multiple sites, there are incompatibilities in the interfaces, in the definition and representation of data, which are often connected to the platform containing them. Furthermore, there are many metadata respect of the information and problems arise in the transmission of data over networks [8] . Moreover, "Big Data Also Brings About New Opportunities for discovering new values, Helps us to gain an in-depth understanding of the hidden values and incurs also new challenges, for example how to effectively organize and manage such datasets" [9].

Advances in information technology (IT), the rapid growth of so-called cloud computing and the Internet of Things (IoT) [9] have increased opportunities to generate and accumulate data, exceeding the capacity of researchers and technology companies to respond to this problem with a systematic and integrated solution. Until 2003, "five exabytes (1018 bytes) of data

were created by human. Today this amount of information is created in two days. Big Data requires a revolutionary step forward from traditional data analysis, characterized by its three main components: variety, velocity, and volume " [10] . "For solutions of permanent storage and management of large-scale datasets disordered, distributed file systems [24] and NoSQL [26] databases are good choices" [9].

McKinsey Global Institute [11] , the potential of Big Data is mainly in five sectors: Healthcare, Public Sector, Retail, Manufacturing and Personal Location Data. Moreover, the research trends in the field of Big Data analytics, point to the heterogeneity of the data and subsequent incongruity that occurs in highly unstructured data; the scalability; the combination of RDBMS and NoSQL Database Query Optimization Systems Issues in HiveQ [7].

Experience with EEG records and derived information concerning this work shows that we are in the presence of a problem of massive data rather than a problem of Big Data, however, they share many of the characteristics and problems associated with big data. In fact, the volume of data not only increases but that each action generates new records or associated metadata that systematically increase storage needs. Moreover, when there is greater availability of instrumental EEG, the growth rate of the volume of data increases linearly in proportion to the increased availability of devices.

Treatment, storage organization and EEG records and their analysis, share the problems inherent of Big Data, in several aspects:

- There are differences between different instruments to capture EEG records and have a high degree of inconsistency.
- The EEG data and its derivatives within the scope of those who is called Big Data Collections [8]. EGG data management is fully compatible with the integration of RDBMS, particularly in the management of metadata to facilitate analytical EEG records.
- EEG records management and its derivatives is a data management problem high scalability. In many types of research, data usually are not discarded but remain stored for future reference, links or reviews.

Although there storage formats for EEG records, such as the European Data Format (EDF), these records are rather unstructured nature. "In order to design meaningful analytics, it is mandatory that big data input sources are transformed into a suitable, structured format" [7].

2 Stages in the treatment of EGG records

The processing and analysis of EEG data go through different stages or states. Initially, after his capture, the EEG records untreated (called RAW) can contain different types of disturbances (artifacts), which should be identified and subtracted from the register, before applying methods of analysis. Then the treatment of previously released data, which will be subjected to different methods, create new records of information (in digital format), which are the input stage of analysis and preparation of scientific reports and clinical report. The different stages being experienced treatment EEG records can be seen in Figure 1.

The following sections provide a brief description of each of the stages or states of the processing and analysis of EEG registrations.

2.1 Data cleaning

The aim the cleaning process is the removal of artifacts [2]. The results of the data cleaning process have important effects on EEG records both in its duration as its size. During the

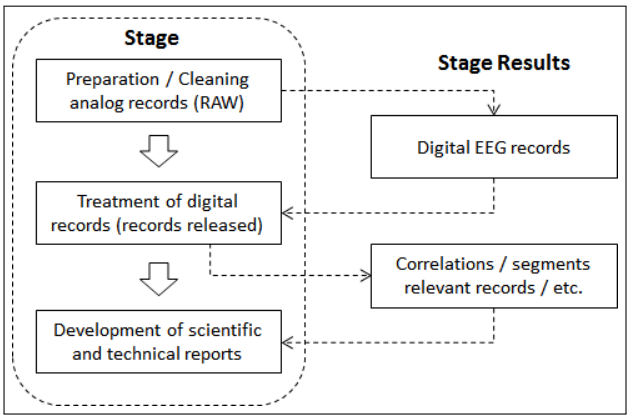


Figure 1: Steges in the treatment of EEG records

process, the analog sequences captured in each of the channels (RAW) are converted to digital streams. Thus, the accuracy and relevance of the digital signal generated will depend on the method parameters that are applied in the purification process.

There are different types of alternative processes for the cleaning of raw records Each EEG record may be subjected to a sequence of various cleaning methods. Some of these results will be discarded during treatment and others will be stored for later analysis. In order to properly interpret the results of the clearance records, you must record the sequence of the methods applied, and the specific values of the parameters that each of the methods was applied.

In the case of EEG records considered in this work, methods and purification processes used are the following:

- Importing Data: Due to the Emotiv EPOC instrument records the sensor signals in 14 useful channels (called: AF3, F7, F3, FC5, T7, P7, O1, O2, P8, T8, FC6, F4, F8 , AF4), a file format of European data format (EDF), the import is done through the built-in tool BIOSIG EEGLAB [4] software. Generally, you must import all channels, but there is the possibility of importing a few.

- Epoch's selection: It refers to select periods of records that contain events of interest to the investigation, in which people objects of study, they are subject to some kind of stimulus.

At this stage, you can also perform other actions considered pre-processing such as changing the sampling rate, filter the data or re-reference them.

2.2 Processing and data exploration

It is understood by data processing, to all actions that aim to inquire into the information contained in the records or derive new information from existing data. There are different methods and techniques that can be used at this stage [4]. The specific order in which they are used, the number of times that it was applied as well as its parameters, depend on the objectives of the research and the findings that researchers performed during the same.

Processing methods according to their purpose and their effect on the data can be classified into:

- Exploratory methods: Specifically, rather than processing methods they are techniques to explore the characteristics and peculiarities of each EEG record. Generally make use of the help of graphical tools that allow the display of one or more channels, thus forming a

better impression and understanding of the records. Including loading and channel display (load and view channel locations), ERP Plotting images, and Plotting Spectra and Maps Channel [4] are frequently mentioned.

- Methods of processing and generation of macro data: It refers to those methods that act on the data to modify or to generate new data. Within this group are, for example, those for determining the baseline (to avoid skewed by the presence of low-frequency artifacts or analysis), ICA using data decomposition [2], work with ICA components, the decomposition time / frequency [4], etc.

3 Context of the capture, storage, processing, and analysis of EEG data

Scientific and clinical studies that require registration, treatment, storage and analysis of EEG data, demand the articulation of investigative processes as well as different types of elements, such as clinical patients, volunteers, premises, equipment, technical staff, clinicians, and researchers (see figure 2). The relevant part of these activities corresponds to drawing conclusions: technical, scientific and clinical diagnostics, which must be properly documented.

The intervention of so many different factors, not only generates abundant records of information of different nature but also requires an organization and storage that includes the different interactions between different components, in order to avoid levels of confusion and distancing from reality, a product the abundant data.

This paper addresses the definition of a data model (see figure 4) that supports the development of an information system based on a storage strategy and organization that considers existing relevant interactions in the context of studies related to the EEG data.

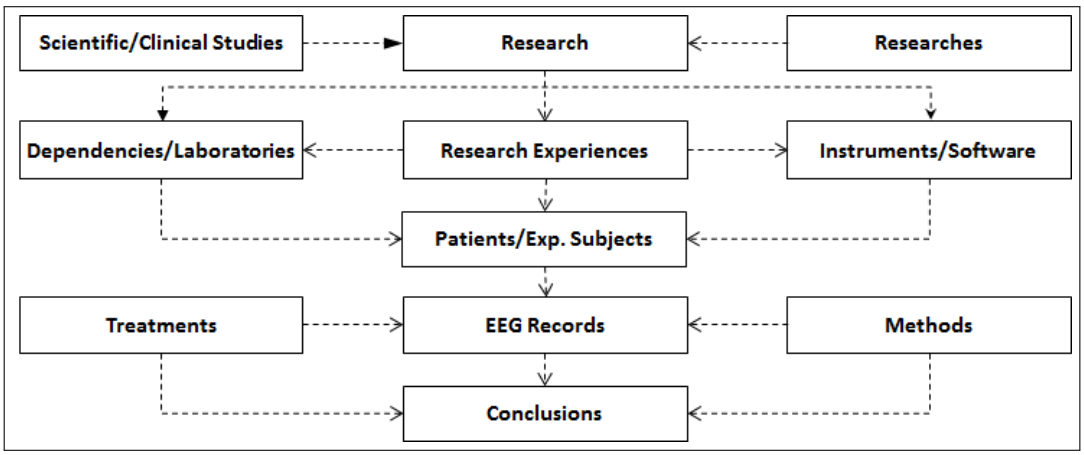


Figure 2: The context in which they exist and the EEG data is generated

The following sections describe the conceptual data model developed to meet the requirements described previously.

3.1 Scientific/clinical studies, research and dependences

In any context of treatment and analysis of EEG data, there will be planned investigative processes (projects) conducted by researchers or clinical staff specialized, which will be assigned to a set of physical units (clinics or laboratories), where all the experimental activities will be

conducted. The enclosures also house the instruments and computers with the necessary software for the treatment of captured EEG records.

3.2 Research experiences and records

The projects require the development of various clinical or experimental activities, which are subject to strict protocols, which therefore must be developed, conducted and supervised by qualified researchers and clinicians. The result of these activities is the generation of several EEG records (RAW), which will be generated in each of the sessions which will be subjected patients or experimental subjects. Therefore, it comes to many EEG records, which require adequately linked to experiences in which they originate, the instruments used, laboratory personnel and dependencies in which they perform.

3.3 Dimensioning and predicting storage needs

The storage, management, processing, and documentation of the results of processing and analysis of EEG records obtained from research and clinical studies, requires a hardware platform that, in addition to processing power, ensure responses in appropriate time lapses and adequate storage management. Investigations that require storage of EEG records and derivates of them require computers able to process and manage files that may have average sizes of hundreds of megabytes.

The type of study that originates the data, affect considerably on the number of records, duration, and size. In the medical field, the amount, duration, and size of EEG records is directly related to the pathology under study and its severity. Experts know beforehand the size and approximate duration of EEG records that are necessary to diagnose with a significant degree of reliability of the presence or absence of certain pathologies.

However, in the field of new applications of EEG technology, which have opened to very different areas outside of classical medicine, the problems have storage requirements and processing needs very variability. Any useful prediction requires the determination of significant variables that serve as the basis for accurate estimates.

Suppose that S is the set of study areas related to EEG data and is I_S the set of research (projects) linked around these areas, then let:

- $P(I)$: The set of all persons involved in a research I, $I \in I_S$.
- $R(x, I)$: The set of EEG records captured a person x in an investigation I, $I \in I_S$ and $x \in P(I)$.
- $RD(x, t, r)$: The set of records derived from treatment of a record EEG r obtained from a person x in research I, $I \in I_S$, $x \in P(I)$, and $r \in R(x, I)$.

If we denote by $I(r)$ to all of the records included in a full investigation I, then the storage space of research related to the different areas of study contained in S, is given by the following expression:

$$Z(S) = \sum_{I \in I_S} \sum_{x \in P(I)} \sum_{r \in R(x,I)} (Z(r) + \sum_{r_d \in R_d(x,I,r)} Z(r_d)) \tag{1}$$

where $Z(x)$ represents the size of an object x measured in MB.

To study this problem, It has been chosen a set of indicators whose behavior is expected to permit characterize the different types of studies and estimating storage needs requiring treatment and analysis of neurophysiological information.

The indicators initially considered are as follows:

Indicator	Description
MB	Amount of storage required by RAW records and derivatives in each investigation.
NS	Amount of subjects involve in each investigation.
NR	Amount of records RAW captured in each investigation.
ND	Amount of (digital) derived records RAW records in each investigation.
MS	The total sum of the minute records captured in a research.
NI	Amount of reports generated in each investigation.

Table 1: Indicators relating to a representative set of EEG investigations.

3.4 Organization of processing methods and reports

One of the ultimate goals of the research is drawing conclusions and scientific / technical reports. This requires the systematic application of various methods of processing and analyzing the results, which will provide researchers with lights on the nature of the phenomena under study.

The effectiveness and relevance of each method are directly related to its parameters and the order or sequence in which they are applied. In many cases, each method may require multiple applications to find a suitable parameterization for the problem, something that must be recorded in the study as a log for further review.

The characterization of a processing method m any can be represented by $m(n, \vec{V})$, where n represents the number of parameters involved in the application of the method and $\vec{V}$ is a vector containing the formal description of parameters.

Processing methods are applied to the original EEG records (RAW) or records arising from the application of previous methods (eg cleaning artifacts). The fact of applying a processing method on a specific EEG record r, can be represented by the cinchona $(m, r, t, n, \vec{v})$, such that the vector $\vec{v}$ contains the specific values of the parameters applied to register r at time t, which can give rise to a new data record $r\prime$.

Thus, the full process a record EEG r can be represented as the set

$$(m(n_i, \vec{V})_i, r_i, t_i, n_i, \vec{v_i})/i =, 1 \ldots k \tag{2}$$

We denote as $M(r)$ a sequence of methods applied in a temporal sequence $t_1, \ldots, t_k$ over an r record.

Given the research process, I involves the generation of one or more records for each individual, we represent by $M_x(r)$ the set containing all sequences $M(r)$ associated with all EEG records applied to the subject S.

Finally, the results obtained from the analysis of all processing sequences contained in $M(r)$ in a research I, give rise to many technical documents or scientific reports that should be properly cataloged and stored for any use present or future.

Scientific and clinical reports contains the conclusions of the analysis made in the investigation and therefore should be linked to each of the records and data used in its construction.

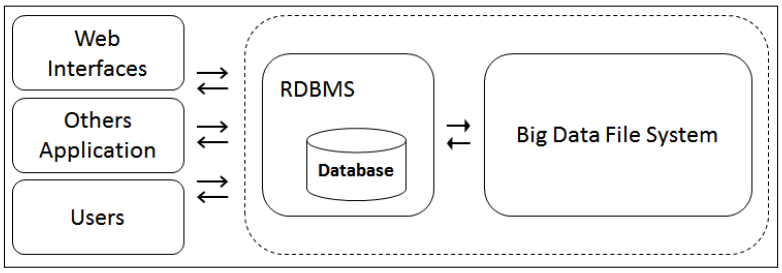


Figure 3: Interaction between system components

4 Data model for the organization, storage and processing of large data sets EEG

It needs to consider the context in which EEG records are recorded, processed and analyzed from multiple investigations; consider that the data from the processing of each of the records are in the presence of a problem of management of complex data, which also requires the coordination and integration of multiple sources of information and technologies. To keep the whole meaning of the information is necessary to include descriptive information concerning research, technical reports, and the conclusions drawn from the records and the various associations that occur between these entities.

The EEGLAB [3] [4] tool has addressed this problem from the standpoint of ease of processing and exploration of the data but does not provide for integration into the information system of the context in which the records are generated. Neither includes information on technical documents with the conclusions drawn from his treatment.

This paper presents a conceptual model for the organization, storage and processing of EEG records (under the Model Entity / Relationship [5]), which takes over the metadata associated with the problem. Considers context information where records are generated. It takes over the need to store the associations between records, treatment, scientific papers and technical reports, which contain the findings of the analysis developed.

This proposal consists in adding information on the context in which data are generated to information concerning the organization, storage and processing of EEG records. This is achieved by a technology-based relational database component, which incorporates contextual information and semantic elements that give meaning to data volumes (see figure 4). Subsequently, on this platform may implement other applications, such as web-based interfaces or other tools that use the metadata system for maintaining and managing the links between different elements of information required to manage information EEG.

4.1 Scientific/clinical studies, research and dependences

The data EEG focus of this study is captured in scientific or clinical studies during the development of properly planned processes (projects), conducted by researchers or clinical professional, in installations or laboratories that house the instruments required to capture data and experimentation. The projects are framed within these studies, and in turn, researchers or clinicians may be linked to different projects. The diagram is shown in Figure 4 depicts this dynamic. In particular, it must be considered clearly states that all projects should be assigned to a field concerning clinical trials or scientific experience, which comprise a wide range of investigations.

4.2 Research, experimental situations, experimental subjects and EEG records

Normally, an investigation involves making different experimental situations or activities, which can obtain new EEG records of experimental subjects, using various instruments. Then each EEG record, although will be associated with only one subject and a specific instrument can be used in other investigations. A similar situation applies to clinical records obtained from patients.

This situation creates a set of EEG records, previously termed I (r), which can be represented as in Figure 4 by record entity.

In research contexts addressed in this work, records obtained from the application of an individual experience, are always linked to a single instrument. However, you can simultaneously record multiple records across different instruments. This situation also has been reflected in Figure 4 as an aggregation (high-level entity) between Experience and Instruments entities.

On the other hand, as shown in Figure 4, the association between a Research and high-level entity Experience/Instruments, is the type many to many, because an investigation can involve more than one pair Experience/Instruments, and in turn, each pair experience/instrument may be associated with multiple investigations. This relationship between Experience/Instrument and Research frequently applied to clinical studies where the same experimental situation, for example, diagnostic procedures are applied to multiple patients. On the other hand, people involved in an investigation or are subjected to a clinical procedure may eventually be required to participate in other research or studies.

4.3 Methods, sequences processing, and parameter settings.

The systematic and successive application of various treatment methods on EEG records in RAW format or those derived from previous treatments is an essential feature of EEG data analysis. Though several treatments do not give useful results, many treatments if they are significant for research, so that should be recorded, including parametrization, identifying the set of input data, and the order and the moment in time they were applied. Consequently, in the exploration and processing of each EEG recording, you must register the application sequence of processing methods. The explorations on the results, all associated with a set of timestamps. Processing said the sequence of previously denoted by the expression (2).

These timestamps introduce a natural chronological order that besides being a historical record of the treatment records, allows us to analyze the strategy analysis and correct it, if necessary. This situation it's represented in Figure 4. The application of different parametrized methods results in the generation of new data records, which add to existing records. This situation is collected as a reflexive relationship called "originates", between the high-level entity called "applies" and the entity called Record.

4.4 Management of technical reports

It is considered that, in general, an investigation involves different experimental situations, which are applied to different people. This creates a set of EEG records, previously has been called $I(r)$.

The processing of EEG records, either through exploratory methods or data processing allows obtaining derived information that is the basis of the conclusions contained in the technical and scientific reports. The preparation of the documents can take information from many records corresponding to a specific research process or multiple investigations, a reason that is it necessary to link the EEG records all the scientific/technical reports involved.

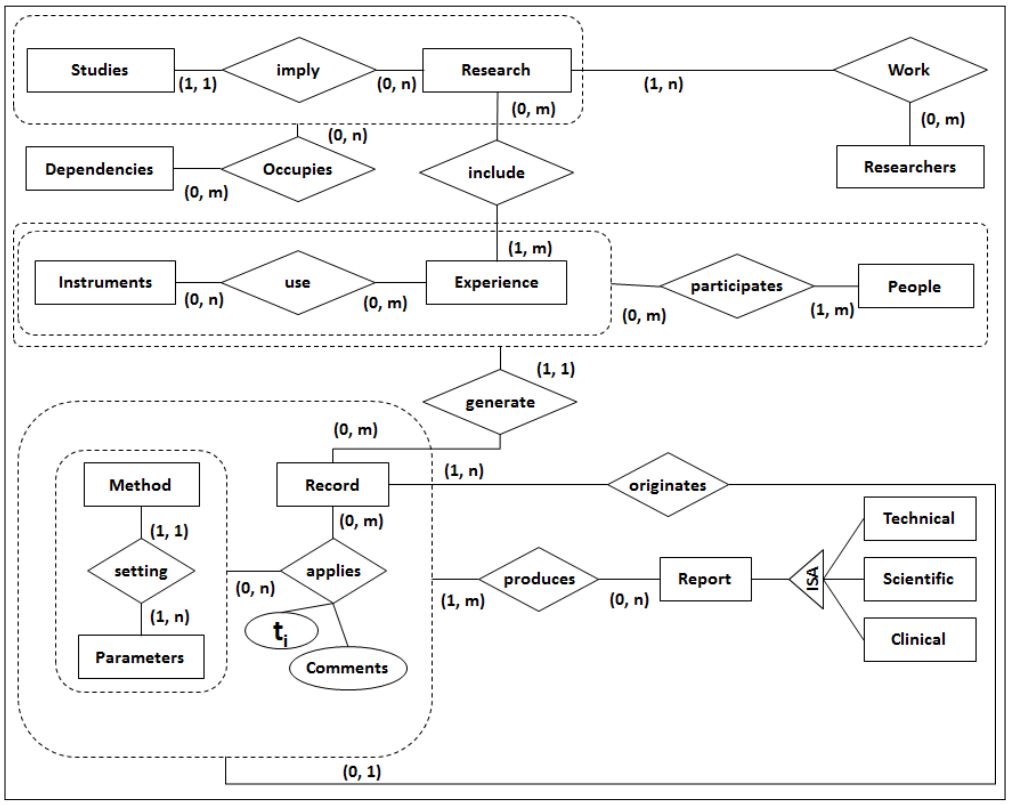


Figure 4: Conceptual model for the context of EEG data

The biggest challenge in the treatment of biometric information is not only the need for storage space for RAW records or derivates but also store reports and other results containing the summary and details of the conclusions.

4.5 Overview of the model for the management, storage and processing of EEG records

Integrating all requirements described in the preceding paragraphs, is represented in the conceptual data model is presented in figure 4. To simplify the figure, the attributes have not been incorporated into the model.

This data model aims to implement a database (with documentary features), to incorporate contextual information and references to the physical location where the data is stored (URL). Specifically, the purpose of the database is as follows: (1) Integrate records and metadata to facilitate the analytic EEG records and derived data; (2) Establish a catalogue of EEG records, scientific and technical documents as well as from the investigation; (3) Allow the generation of a system capable of recording the activities of research centres and evolve if necessary; and (4) Allow the development of tools for analysis and processing of data, which are made in the context of the system, automatically integrating partial and final results of interest to research centres.

To facilitate the implementation of the de model in any relational platform and therefore the portability of applications that make use of it, in figure 5 shows the standard logical design (DLS) of the conceptual model.

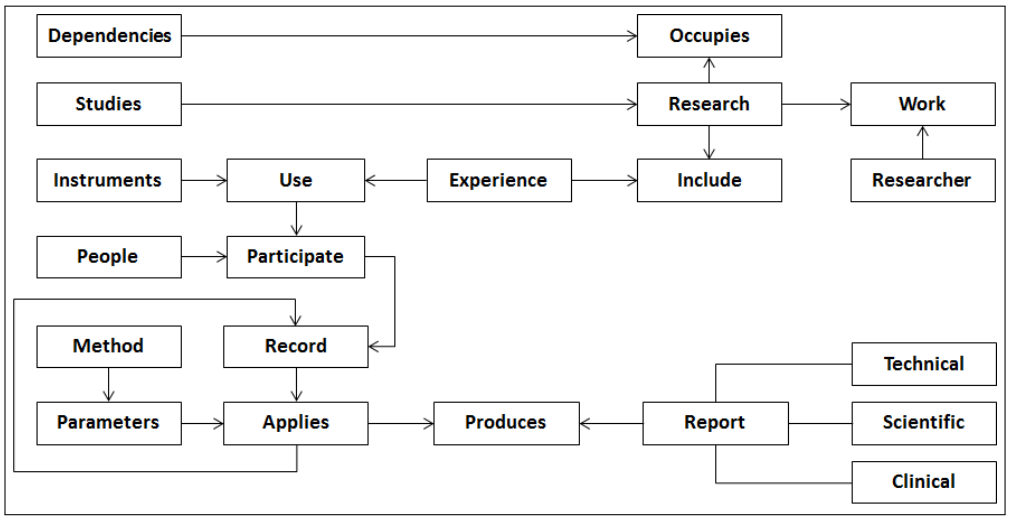


Figure 5: DLS for the conceptual model about context of EEG data

Conclusions

The growing needs to store large amounts of information relating to various areas of human and scientific work, has prompted different approaches and approaches related to the storage of information, is now called Big Data. Although these proposals address many of the problems associated with the storage, access, and distribution of information on computer media, they do not have sufficient tools to manage the semantic elements of the stored information, which is specific to each problem type. In the case of EEG records and technical and scientific reports, it is required of mechanisms for the organization, storage and processing of data, allowing the integration of the various elements of information contained in the records and mainly those generated from their treatment.

Current developments in the area of storage [13] [14] provide an alternative for managing large volumes distributed in different storage media, which however are limited by the heterogeneity of the different types of data. Proposals made in this document provide a solution for these semantic constraints in the context of specific data, by integrating other technologies, such as database services (DaaS) [15] [7], which also allow manipulation of metadata and integrate different types of information, enabling the development of applications and interfaces that automate the recording and metadata generation. In addition, the technology database allows inheriting mechanisms to ensure the confidentiality of information regarding experimental subjects, the analysis, and conclusions of the records, research, and researchers.

In this scenario the data model proposed is able to leverage solutions in the field of Big Data, incorporating semantic elements that enable researchers and technical staff not to lose control over data and new knowledge derived from them. The data model becomes a significant interface between the information processing methods, facilitating data analysis and generation of conclusions. Moreover, the increased availability of metadata and recording timestamps on different moments related to the processing, analyzing and drawing conclusions, promote more efficient use of computing resources by decreasing the need to reprocess data unnecessarily when those results are registered for the system level.

Finally, the use of technologies of conventional databases available in the field of free software or proprietary contributes to the greater integration of information value in the new scenario underlying the current developments in the field of human-computer interfaces in science and medicine.

Acknowledgment

This work was developed under the project "Definición, desarrollo e implementación de nuevos indicadores y procedimientos para la evaluación y seguimiento del rendimiento académico estudiantil y el desempeño docente, en base información curricular, variables psico-neuro-cognitivas y herramientas basadas en matemática y lógica difusa", which receives funding from the Department of Scientific and Technological Research (DICYT) of the University of Santiago of Chile (USACH), Chile.

Bibliography

[1] European Data Format (EDF). [En línea], Disponible at `http://www.edfplus.info/` [Accesed: 17-nov-2015].

[2] T.-P. Jung et al. (1998); Removing electroencephalographic artifacts: comparison between ICA and PCA, *Neural Networks for Signal Processing VIII, 1998. Proceedings of the 1998 IEEE Signal Processing Society Workshop*, 63-72, DOI: 10.1109/NNSP.1998.710633

[3] A. Delorme, S. Makeig (2004); EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis, *Journal of neuroscience methods*, 134(1): 9-21, 2004.

[4] EEGLAB Tutorial: Table of Contents, `http://sccn.ucsd.edu/eeglab/eeglabtut.html`

[5] Chen Peter (1976), The Entity-Relationship Model - Toward a Unified View of Data, *ACM Transactions on Database Systems*, 1(1): 9-36.

[6] E.F. Codd (1970); A relational model of data for large shared data banks, *Communications of the ACM*, 13(6):377-387.

[7] A. Cuzzocrea, I.-Y. Song, y K. C. Davis (2011); Analytics over large-scale multidimensional data: the big data revolution, *Proceedings of the ACM 14th international workshop on Data Warehousing and OLAP*, 101-104.

[8] M. Cox, D. Ellsworth (1997); Managing big data for scientific visualization, *ACM Siggraph*, 97: 21-38.

[9] M. Chen, S. Mao, Y. Liu (2014); Big Data: A Survey, *Mobile Networks and Applications*, 19(2): 171-209.

[10] S. Sagiroglum, D. Sinanc (2013); Big data: A review, in Collaboration Technologies and Systems (CTS), *2013 International Conference on*, 42-47.

[11] J. Manyika, M. Chui, B. Brown, J. Bughin, R. Dobbs, C. Roxburgh and A.H. Byers (2011); *Big data: The next frontier for innovation, competition, and productivity*, McKinsey Global Institute, 2011.

[12] S. Kaisler, F. Armour, J. A. Espinosa, W. Money (2013); Big Data: Issues and Challenges Moving Forward, *2013 46th Hawaii International Conference on System Sciences*, 995-1004.

[13] Zikopolous Paul, Deroos Dirk, Deutsch Tom, Lapis George (2012); *Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data*, McGraw-Hill, 2012.

[14] HPCC Systems. [Online]. `https://hpccsystems.com/`

[15] H. Hacigumus, B. Iyer, S. Mehrotra (2002); Providing database as a service, en Data Engineering, *2002 Proceedings. 18th International Conference on*, 29-38.

[16] B.B. Biswal et al. (2010); Toward discovery science of human brain function, *Proceedings of the National Academy of Sciences*, 107(10): 4734-4739.

[17] A. Paul Alivisatos et al. (2012); The Brain Activity Map Project and the Challenge of Functional Connectomics, *Neuron* 74(6): 970-974, DOI: `http://dx.doi.org/10.1016/j.neuron.2012.06.006`

[18] Monti M. M., Parsons L. M., Osherson D. N. (2009): The boundaries of language and thought in deductive inference, *Proceedings of the National Academy of Sciences*, 106(30): 12554-12559.

[19] Evan Heit (2015); Brain imaging, forward inference, and theories of reasoning, *rontiers in Human Neuroscience*, January 2015, Vol. 8, Article1056.

[20] Vinod Goel, Brian Gold, Shitij Kapur (1997); Sylvain Houle, The seats of reason? An imaging study of deductive and inductive reasoning, *NeuroReport*, 8: 1305-1310.

[21] Ranjit A. Thuraisingham, Georg A. Gottwald, On multiscale entropy analysis for physiological data, Physica A: Statistical Mechanics and its Applications, Volume 366, 1 July 2006, Pages 323-332.

[22] Hass L.F. (2003); Hand Berger (1873-1941), Richard Caton (1842-1926), and electroencephalography, *Journal of Neurology, Neurosurgery and Psychiatry*, 74(1):9, doi:10.1136/jnnp.74.1.9.

[23] Kececi H., Degirmenci Y. (2008); Quantitative EEG and cognitive evoked potentials in anemia, *Clinical Neurophysiology*, 38(2): 137-143, doi:10.1016/j.neucli.2008.01.004

[24] Rami N. Khushaba et al. (2013); Consumer neuroscience: Assessing the brain response to marketing stimuli using electroencephalogram (EEG) and eye tracking, *Expert Systems with Applications*, 40:803-3812.

[25] Fredi E. Palominos, Hernan Diaz, Felisa M. Cordova, Cañete Lucio, Claudia A. Duran, Model for the organization, storage and processing of large data banks of physiological variables, *Computers Communications and Control (ICCCC), 2016 6th International Conference on*, e-ISBN 978-1-5090-1735-5, IEEE Xplore,173-179, DOI: 10.1109/ICCCC.2016.7496757

[26] Díaz, Hernán, L. Cañete, F. Palominos, C. Costa y F. Córdova (2012); Neurotechnologies for Education Improvement: Self-Knowledge After Opening the Black-box, *Conference Proceedings of the International Symposium Research and Education in Innovation Era. ISREIE, 4th Edition Arad. Journal Plus Education*, 8(2): 44 - 52.

[27] Paul Howard-Jones et al. (2013); Neuroscience and Education: Issues and Opportunities, TLRP Teaching and Learning Research Program; ESRC Economic and Social Research Program, University of London, ISBN: 0-85473-741-3, 2013.

[28] Maria Kozhevnikov (2007); Cognitive Styles in the Context of Modern Psychology: Toward an Integrated Framework of Cognitive Style, *Psychological Bulletin*, 133(3): 464-481.

[29] Matthew D. Lieberman (2007); Social Cognitive Neuroscience: A Review of Core Processes, *Annu. Rev. Psychol.* 58:259-289.

[30] Raja Parasuraman, George Mason, Glenn F. Wilson (2008); Putting the Brain to Work: Neuroergonomics Past, Present, and Future, *Human Factors*, 50(3): 468-474. DOI 10.1518/001872008X288349.

Author index