# DSIE'17

## Proceedings of the 12th Doctoral Symposium in Informatics Engineering

### Editors:

A. Augusto de Sousa

Eugénio Oliveira

January 31st, 2017
Porto, Portugal

# COPYRIGHT

1ª Edição/ 1st Edition 2017
ISBN: 978-972-752-207-1
Editors: A. Augusto Sousa and Eugénio Oliveira
Proceedings Design: Pedro Santos
Graphical Design/Website: José Macedo, Massanobu Tachikawa and Ricardo Lopes
Faculdade de Engenharia da Universidade do Porto
Rua Dr. Roberto Frias, 4200-465 Porto

DSIE'17 SECRETARIAT:
Faculdade de Engenharia da Universidade do Porto
Rua Dr. Roberto Frias, s/n
4200-465 Porto, Portugal
Telephone: +351 22 508 21 34
Fax: +351 22 508 14 43

E-mail: dsie2017@fe.up.pt
Symposium Website: www.fe.up.pt/dsie2017/

# FOREWORD

*STEERING COMMITTEE*

These are great times for those who have chosen to pursue advanced studies in Computer Science and Informatics Engineering. Modern world is literally being shaped by computers and, above all, by flexible, adaptable, evolving, friendly, secure and intelligent software.

DSIE - Doctoral Symposium in Informatics Engineering, now in its 12th Edition, is an opportunity for the PhD students of the FEUP Doctoral Program in Informatics Engineering (ProDEI) to show up and prove they are ready for starting their respective theses work.

DSIE meetings have been held since the scholar year 2005/06 and the main goal has always been to provide a forum for discussion on, and demonstration of, the practical application of a variety of scientific and technological research issues, particularly in the context of information technology, computer science and computer engineering. DSIE symposium comes out as a natural conclusion of mandatory ProDEI course called "Methodologies for Scientific Research" (MSR) leading to a formal assessment of the PhD students first year's learned competencies on those methodologies.

The above mentioned specific course (MSR) aims at giving students the opportunity to learn the processes, methodologies and best practices related to scientific research, particularly in the referred areas, as well as to improve their own capability to produce adequate scientific texts and articles. With a mixed format based on a few theoretical lessons on the meaning of a scientific approach to knowledge, together with multidisciplinary seminars and tutorials, the course culminates with the realization of the DSIE meeting. DSIE is seen as a kind of laboratory test for the concepts learned by students. In the scope of DSIE, students are expected to simultaneously play various roles, such as authors of the submitted articles, members of both scientific and organization committees, as well as reviewers, duly guided by senior lecturers and professors.

DSIE event is then seen as the opportunity for the students to be exposed to all facets of a scientific meeting associated with relevant research activities in the above mentioned areas. Although still at an embryonic stage, and despite some of the papers still lack of maturity, we already can find some interesting research work or promising perspectives about future work. At this moment, it is not yet essential, nor often possible, for most of the students in the first semester of their PhD, to produce sound and deep research results. However, we hope that the basic requirements for publishing an acceptable scientific paper have been fulfilled.

Each year DSIE Proceedings include papers addressing different topics according to the current students' interests in Informatics Engineering. This year, the tendency is on Data and Text Mining, Agent-based Computing and network analysis, Frameworks and Security, Software Engineering, Multimedia, Cloud and Databases services.

The complete DSIE'17 meeting lasts one day and includes also one invited talk by an outstanding researcher in Streaming Data Mining.

Professors responsible for ProDEI program's current edition, are proud to participate in DSIE'17 meeting and would like to acknowledge all the students who have been deeply involved in the success of this event that, hopefully, will contribute for a better understanding of the themes addressed during the above referred course, the best scientific research methods and the good practices for writing scientific papers and conveying novel ideas.

***Porto, January 2017***

***Eugénio Oliveira and Augusto Sousa (ProDEI)***

# FOREWORD

*ORGANIZING AND SCIENTIFIC COMMITTEES*

The organization and scientific committees are glad to welcome you to the Doctoral Symposium in Informatics Engineering, DSIE'17. The symposium is an important mark to the students beginning their doctoral journey. It provides an excellent opportunity to improve their scientific research capabilities as well as leverage the feedback from field experts.

In this 12th edition we extend the scope of the symposium, holding for the first time in DSIE a poster session specially designed for master's students wishing to present their ongoing thesis projects. Accepted papers address several important topics from informatics engineering, such as data mining, intelligent systems or software testing. Moreover, they cover important social and industrial issues, such as healthcare and transportation systems. As Keynote Speaker, professor Dr. João Gama, a world class researcher, talks about the increasing impact data mining has in our society.

We would like to acknowledge everyone that was part of the organization of the event. The Steering Committee that guided us through every detail of the organization. The Senior Scientific Committee for providing insightful ideas about the research lines of the students. Professor João Gama for accepting our invitation as Keynote Speaker. The secretariate of the Department of Informatics Engineering, namely Sandra Reis, whose support in the organization was fundamental. Finally, we thank our sponsors, without them we wouldn't be able to organize this event.

***Porto, January 2017***

***Guilherme Soares and Vitor Cerqueira (Organization Committee Chair)***

***Gudilak Constantino and Maria Pedroto (Scientific Committee Chairs)***

# CONFERENCE COMMITTEES

## STEERING COMMITTEE

A. Augusto Sousa

Eugénio Oliveira

## ORGANIZING COMMITTEE CHAIR

Guilherme Soares

Vitor Cerqueira

## Organizing COMMITTEE

Denis Ogura

Gil Rocha

Gudilak Constantino

José Macedo

Maria Pedroto

Massanobu Tachikawa

Michael A.P. Domingues

Pedro Santos

Ricardo Lopes

Thiago Baldês

Tiago Soares Costa

## SCIENTIFIC COMMITTEE CO-CHAIRS

Gudilak Constantino

Maria Pedroto

## SENIOR SCIENTIFIC COMMITTEE

Alípio Jorge

Ana Paiva

Ana Paula Rocha

Carlos Soares

Fábio Pinto

Filipe Coelho

Henrique Lopes Cardoso

João Barbosa

João Pascoal Faria

João Mendes Moreira

Rosaldo Rossetti

Rui Maranhão de Abreu

Rui Rodrigues

Sérgio Sobral Nunes

## JUNIOR SCIENTIFIC COMMITTEE

Denis Ogura

Gil Rocha

Guilherme Soares

José Macedo

Massanobu Tachikawa

Michael A.P. Domingues

Pedro Santos

Ricardo Lopes

Thiago Baldês

Tiago Soares Costa

Vitor Cerqueira

DSIE'17 – Doctoral Symposium in Informatics Engineering is sponsored by:

## Platinum Sponsor

YAZAKI

## Silver Sponsor

ARMIS GROUP

## Supporters

PORTO

CRITICAL SOFTWARE

mediagaps
Bridging your workflow Gaps!

SIEMENS Healthineers

## Institutional

U.PORTO

U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

DEI Departamento de Engenharia Informática

PRODEI | Programa Doutoral em Engenharia Informática

# CONTENTS

## INVITED SPEAKER

## POSTER SESSION

## SESSION 1 - SOFTWARE TESTING

## SESSION 2 - SIMULATION AND OPTIMIZATION

SESSION 3 - INTELLIGENT SYSTEMS

SESSION 4 - MACHINE LEARNING & DATA MINING

# INVITED SPEAKER

## PROFESSOR JOÃO GAMA

João Gama is Associate Professor of the Faculty of Economy, University of Porto. He is a researcher and vice-director of LIAAD, a group belonging to INESC TEC. He got the PhD degree from the University of Porto, in 2000. He has worked in several National and European projects on Incremental and Adaptive learning systems, Ubiquitous Knowledge Discovery, Learning from Massive, and Structured Data, etc.

He served as Co-Program chair of ECML'2005, DS'2009, ADMA'2009, IDA' 2011, ECML/PKDD'2015, EPIA 2017, and DSAA 2017. He served as track chair on Data Streams with ACM SAC from 2007 till 2016. He organized a series of Workshops on Knowledge Discovery from Data Streams with ECML/PKDD, and Knowledge Discovery from Sensor Data with ACM SIGKDD.

He is author of several books in Data Mining (in Portuguese) and authored a monograph on Knowledge Discovery from Data Streams. He authored more than 250 peer-reviewed papers in areas related to machine learning, data mining, and data streams.

He is a member of the editorial board of international journals ML, DMKD, TKDE, IDA, NGC, and KAIS. He supervised more than 12 PhD students and 50 Msc students.

## INVITED TALK

### Datamining For The XXI Century

**Abstract:** Nowadays, there are applications in which the data are modeled best not as persistent tables, but rather as transient data streams. In this talk, we discuss the limitations of current machine learning and data mining algorithms. We discuss the fundamental issues in learning in dynamic environments like continuously maintain learning models that evolve over time, learning and forgetting, concept drift and change detection. Data streams produce a huge amount of data that introduce new constraints in the design of learning algorithms: limited computational resources in terms of memory, cpu power, and communication bandwidth. We present some illustrative algorithms, designed to taking these constrains into account, for decision-tree learning, hierarchical clustering and frequent pattern mining. We identify the main issues and current challenges that emerge in learning from data streams that open research lines for further developments.

# POSTER SESSION

**HealthTalks - How to Improve Health Communication and Personal Information Management**
*João M. Monteiro and Carla Teixeira Lopes*

# HealthTalks - How to Improve Health Communication and Personal Information Management

João M. Monteiro[1] and Carla Teixeira Lopes[1,2]

[1]Faculdade de Engenharia da Universidade do Porto, Porto, Portugal
[2]Instituto de Engenharia de Sistemas e Computadores, Tecnologia e Ciência, Porto, Portugal
{ei11055,ctl}@fe.up.pt

**Keywords:** Health literacy · Self-care · Physician-patient relationship · Personal information management · Speech-to-text · Mobile application

## 1 The Power of Health Information

Health literacy measures the degree to which people can understand their own health situation, what influences it, how to place it into context and how to address it when problems arise in their everyday life [4, 6, 12]. It is of the utmost importance for an individual's well-being, since a high level of health literacy is associated with an improvement of their health condition and reduced costs of healthcare [5]. In Portugal, for example, 49% of people had a "problematic" or "inadequate" level of health literacy in 2014, being that older people and people with lower levels of education presented lower levels of health literacy [1]. In the U.S.A, only 12% of the adults had proficient health literacy in 2003, and over 1 in every 3 adults had a "basic" or "below basic" level [8].

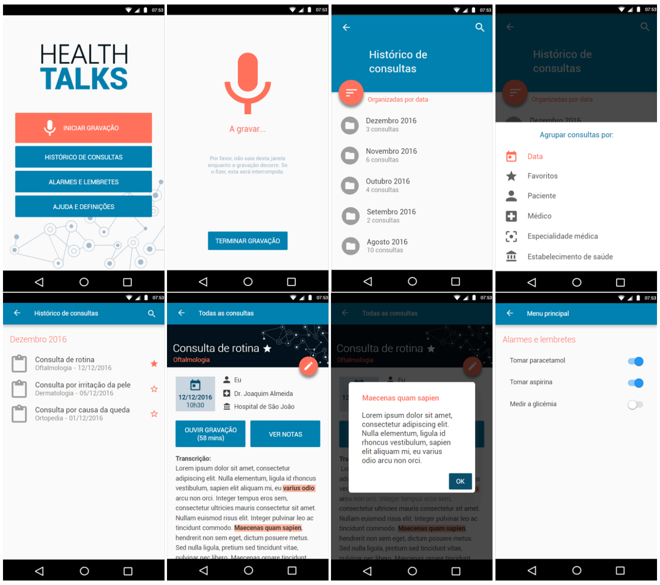
Studies have shown that health professionals have an important role when conveying information, since their communication skills and the complexity of the vocabulary they use may interfere with the patient's ability to understand the message [9, 10]. Obviously, low levels of health literacy exacerbate this type of difficulties. Not understanding the information provided by medical professionals might have a direct impact on people's health [11].

Patients seldom ask questions during consultations, even though more active patients perceive more information [3]. They choose to look for answers online instead: 59% of Internet users in the European Union searched for health information online in 2013 [13], while 80% of American Internet users, or 59% of all American adults, did the same in 2011 [2]. But difficulties in online search usually arise due to the differences between the terminologies used by health consumers and health professionals [13].

All in all, there are several factors that may limit and create imbalances on the patient's access to health information [3, 2]. To add to that, patients may not know if they should save health information they come across and if so, how and where to keep it. When they do take notes, the information they gather becomes fragmented (by location, device, form, and organizations involved), which may difficult the process of looking for it afterwards. Also, the sheer volume of data they find and the vocabulary used may be distressful for them [11].

## 2 How to Empower the Patients

We envision a mobile application, HealthTalks, that will contribute to reducing the inequalities in access to health information and give patients a tool to help managing it better. It will be an Android[1] app that the patient or their guardian can use during a consultation with a physician. The user can choose to record all or only part of the conversation and this will trigger a new consultation entry. Those entries have editable fields (which family member was the patient, where they were, who was the doctor, and what was their speciality) so that they can be easily managed. The user can also write notes on each consultation and mark favourites. When connected to the Internet, they can transcribe the consultation and, in the transcription, they will be able to check definitions for medical terms. Other functionalities such as setting reminders are being considered. Mockups for the app can be seen in Fig. 1.



Por favor, não saia desta janela enquanto a gravação decorre. Se o fizer, esta será interrompida.

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Nulla elementum, ligula id rhoncus vestibulum, sapien elit aliquam mi, eu varius odio arcu non orci. Integer tempus eros sem, consectetur ultricies mauris consectetur sit amet. Nullam euismod risus elit. Integer pulvinar leo ac tincidunt commodo. Maecenas quam sapien, hendrerit non sem eget, dictum posuere metus. Sed nulla ligula, pretium sed tincidunt vitae, pulvinar nec libero. Maecenas ornare tincidunt

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Nulla elementum, ligula id rhoncus vestibulum, sapien elit aliquam mi, eu varius odio arcu non orci.

**Fig. 1.** Initial mockups for HealthTalks (in Portuguese). From left to right and top to bottom: landing screen, recording screen, recordings directory, options to sort the recordings directory, recordings folder, recording details page, recording details page with the definition of a highlighted medical term, and reminders page.

There are several apps available for managing health information (such as MyFitnessPal, MyChart, BG Monitor Diabetes, Pillboxie), for taking notes using speech-to-text (like Evernote, ListNote, SpeechNotes, Google Keyboard), and for giving health

[1] The selection of this specific operating system was based on a) its ubiquity nowadays; b) the familiarity of the authors with its development tools; and c) the easier access by the authors to Android smartphones for testing.

information to patients (e.g.: Patient Journey App, Patient.co.uk, eMed.pt). However, none gathers in itself all the functionalities HealthTalks will have, especially not directed at a Portuguese public, with a dictionary of medical jargon to be recognised in the speech, and a database of medical terms to define. Information management apps such as MyFitnessPal have been found effective as a passive learning device due to the extra information automatically given in an unobtrusive way [7]. That is the same approach we will use in HealthTalks.

## 3 The Road Ahead

The app will use a speech-to-text API in order to transcribe the recordings. After comparing the characteristics of several different speech-to-text programs, Google Cloud Speech API and Microsoft Speech Platform were selected for their customisability and general support. The next step will be to compare them in simulated scenarios in order to choose the one to be used in the final solution.

The application is still in its first phase of development, but the high-level plan can be seen in Fig. 2. Questionnaires are currently being distributed online[2] and are intended to be handed out in Porto's São João Hospital Center as well in order to get a better characterisation of the physician-patient relationship, the degree to which individuals manage their health information and how receptive they would be to a platform such as HealthTalks.

After the results of the surveys are analysed and a speech-to-text API is chosen, the implementation of the app can begin. Future studies will focus on the quality of the transcriptions by comparing different languages, dictionaries and scenarios.

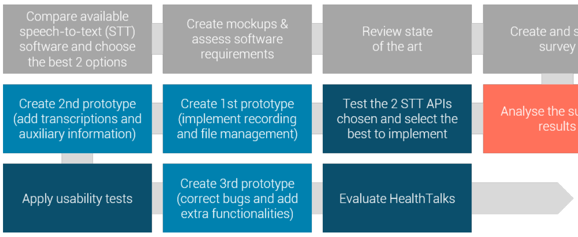

Compare available speech-to-text (STT) software and choose the best 2 options
Create mockups & assess software requirements
Review state of the art
Create and share survey
Analyse the survey's results
Test the 2 STT APIs chosen and select the best to implement
Create 1st prototype (implement recording and file management)
Create 2nd prototype (add transcriptions and auxiliary information)
Apply usability tests
Create 3rd prototype (correct bugs and add extra functionalities)
Evaluate HealthTalks

**Fig. 2.** Expected work plan. Gray items have been completed, orange is currently being done, blue is scheduled to be done in the future. Lighter blue indicates prototypes.

[2] Available exclusively in Portuguese at `https://goo.gl/forms/DgLEoWFOwD4ltLx03`

## References

1. Espanha, R., Ávila, P., Mendes, R.V.: Literacia em Saúde em Portugal - Relatório Síntese. Tech. rep., Fundação Calouste Gulbenkian & Centro de Investigação e Estudos de Sociologia/IUL (2015), `https://gulbenkian.pt/wp-content/uploads/2016/05/PGISVersCurtaFCB\_FINAL2016.pdf` [Accessed 25 January 2017]
2. Fox, S.: The Social Life of Health Information. Tech. rep., Pew Research Center's Internet & American Life Project, Washington, D.C. (2011)
3. Harrington, J., Noble, L.M., Newman, S.P.: Improving patients' communication with doctors: a systematic review of intervention studies. Patient Education and Counseling 52(1), 7-16 (jan 2004)
4. Heijmans, M., Uiters, E., Rose, T., Hofstede, J., Devillé, W., van der Heide, I., Boshuisen, H., Rademakers, J.: Study on sound evidence for a better understanding of health literacy in the European Union. Tech. rep., Consumers, Health and Food Executive Agency (Chafea), Luxembourg (2015)
5. Hur, I., Lee, R., Schmidt, J.: How Healthcare Technology Shapes Health Literacy? A Systematic Review. In: AMCIS 2015 Proceedings (2015)
6. Kutner, M., Greenberg, E., Jin, Y., Paulsen, C.: The Health Literacy of America's Adults: Results From the 2003 National Assessment of Adult Literacy. Tech. rep., National Center for Education Statistics, Washington, DC (2006)
7. Luna, A.L., Guilhermino, B.d.S., Fragoso, K.L., Medeiros, G.L.S., da Rocha, V.C.d.L., da Silva, I.B.A.: O aplicativo MyFitnessPal como recurso promotor da aprendizagem de biologia. Revista Interdisciplinar de Tecnologias e Educação 1(0), 259-262 (2015)
8. Office of Disease Prevention and Health Promotion: America's Health Literacy: Why We Need Accessible Health Information (2008), `https://health.gov/communication/literacy/issuebrief/` [Accessed 9 January 2017]
9. Office of Disease Prevention and Health Promotion: Health Literacy Basics (2016), `https://health.gov/communication/literacy/quickguide/factsbasic.htm` [Accessed 28 November 2016]
10. Osborne, H.: Health Literacy From A to Z: Practical Ways to Communicate Your Health Message. Jones & Bartlett Learning, Burlington, MA, second edn. (2013)
11. Pratt, W., Unruh, K., Civan, A., Skeels, M.: Personal Health Information Management. Communications of the ACM 49(1) (2006)
12. Sørensen, K., Van den Broucke, S., Fullam, J., Doyle, G., Pelikan, J., Slonska, Z., Brand, H.: Health literacy and public health: A systematic review and integration of definitions and models. BMC Public Health 12(1), 80 (2012), `http://bmcpublichealth.biomedcentral.com/articles/10.1186/1471-2458-12-80` [Accessed 22 November 2016]
13. de Sousa, H.M.R.: HealthTranslator: automatic annotation of Web documents in order to assist health consumer's searches. M.sc. thesis, Universidade do Porto (2016)

# SESSION 1

## SOFTWARE TESTING

**Performance testing of open-source HTTP web frameworks in an API**
*Michael A.P. Domingues*

**A Survey on Mutation Testing for Database Applications**
*Pedro Santos*

**Solutions for Automated Mobile Testing Cloud Service**
*Ricardo Lopes*

**Common Threats in Security Testing with Scrum Framework: A Systematic Literature Review**
*Thiago Baldês*

# Performance testing of open-source HTTP web frameworks in an API

Michael A.P. Domingues

Faculty of Engineering, University of Porto, Porto, Portugal
michael.domingues@fe.up.pt

**Abstract.** Upon each day there are more and more services running on the web and we are trying to make to most out of them. These Web Applications (WAs) represent thousands of type of services we can interact with like weather forecast and Application Programming Interface (API) which use by themselves other WAs. API provides at a certain authorization level access to multi-source data using several request endpoints in a multi-tier architecture where performance is critical and it must be ensured the best throughput, response time and availability. This study presents the implementation of three open-source web frameworks (Revel, Gin and Echo) in a corporate API developed by Movvo® (www.movvo.com) in order to evaluate the overall architecture using Apache Benchmark as a load testing tool. Results show Gin contributed to the fastest response times for a set of requests that vary on processing and retrieved data complexity.

**Keywords:** Performance testing, API, Web frameworks, Apache benchmark

## 1 Introduction

Web-services and cloud computing are known concepts in our daily life. Al- most every service we use either in the smartphones or computers have an online architecture that exploits the most from an IT infra-structure to process and execute tasks. The heterogeneity of this technology is ideal to host and deploy services in a flexible way inter or intra-organizational. Its advantages through reduced development time and cost considering a reusability mechanism become significant marks when deploying features and data [1].

Therefore the IT infrastructure performance is crucial and needs to be reliable to respond in stressful conditions with minor failure rates. If it delivers the best performance without redundancy (e.g. backup nodes) it means the system is robust with solid quality testing standards. This represents Quality of Service (QoS) measures and it helps to understand how well web-based applications meet the expected performance through several measures evaluations as availability, response time, load testing and so on.

An API is a service (WA) extremely used across all world by weather forecast agencies, government, browsers and others to serve a set of procedures to access features or data of a specific system or application. Using an API induces the need to use available routes and configurations ensuring there is no other way to manipulate and mutate the data that is retrieved. Ensuring an API effectiveness in the majority of testing methods addresses the same QoS measures as for any web service we know.

The performance of WAs resides mainly in the quality of its architecture. Components that are not ready to be scaled, that get stressed with traffic in- crease, with high latency (disk usage or CPU utilization) are some of the drifts WAs suffer when doing performance testing. It is important to adopt and implement a performance risk management in all phases [2].

When increasing the number of users using the application is expected a peak workload period and when availability and efficiency are not in ideal conditions the system will be deteriorated. This is the time to balance all factors improving hardware resources and implementing an efficient business logic layer including features as caching (see Figure 1).

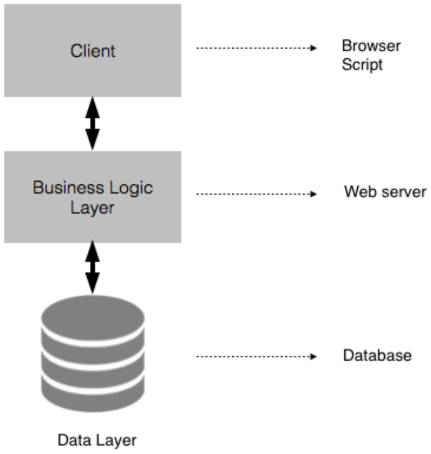


**Fig. 1:** Web Application Architecture

The middle tier also called Web server handles all business logic and serves data to whom consumes it. Major architecture improvements are made mainly in this layer resulting in fast and accurate access to the data storage tier and consequently improved services. But in order to test this type of architecture is important to test individually each tier in distinct web aspects like performance. An interesting measure of testing is load testing which using processes, measures a realistic load based on service parameters and on usage conditions. This called load generator can emulate thousand of concurrent users and it is expected to mimic real user behavior under circumstances of different sessions management [3].

The key use case of this technique is when we want to anticipate an intense traffic or to test the database load dependent on the type of queries done. This set includes also scenarios when a new feature is added and we want to know how the addition is going to affect the response time [3].

To effectively do a load test is important to know what parameters to vary in the service we want to test. Varying the number of virtual users (simulated users accessing the tested application simultaneously) until we achieve the desired load is one possibility but it is essential to test this premise with numerous test values. Evaluating manually the application's workload performance is a burden process wasting too much time but in the last couple of years several testing tools have been developed to suppress this cost [1, 4].

- Apache JMeter
- Apache Benchmark
- NeoLoad
- LoadRunner
- LoadUI
- WebLOAD
- WAPT
- Rational Performance Tester
- Testing Anywhere
- QEngine (ManageEngine)
- WebInject
- WCF Storm
- SoapUI Pro

All of these tools have the purpose of reducing the manual cost and the amount of time needed offering advanced testing capabilities. They all have similar functionalities but Apache JMeter is one of the most complete tool with exploratory performance tests both on static and dynamic resources (files, Servlets, FTP servers, Databases, etc...). LoadUI is the most flexible and interactive test- ing tool with advanced analysis, report generating features and data injection while the application is being tested.

Apache Benchmark on the other hand is a simple HTTP benchmark tool that permits to analyze how many requests per seconds a service is capable of serving. All the configurations are parameterized through command-line flags and even tough it shouldn't be used as the main testing tool (the more tools and results the better) it gives a simple and solid idea of how well the service is running.

This study presents a load testing analysis using Apache Benchmark for a corporate API considering three different web frameworks (Revel, Gin and Echo) that manage the HTTP layer in divergent ways. Results were collected from a set of requests each with complex logic operations and different data structure retrieval. Switching the native HTTP layer with those frameworks each with claimed advantages, allowed us to understand how the API performed with each implementation under stressful testing conditions and with the same data tier.

The organization of this paper consists of following sections: Section 1 lays the basic of this study, Section 2 provides an overview of the state of the art, Section 3 explains the approach and the chosen configurations, Section 4 gives the results and discussion and Section 5 concludes the study along with scope in Section 6.

## 2 Related work

The availability of a pool of configurable computing resources (e.g: services and applications) through a solid, fast and on-demand network is not always granted. The provisioning and the maintenance efforts over-kill the expectations and therefore it is important to have a measure for this cloud computing solutions.

Problems inherent to infra-structures scaling and services migration which are common tasks to any IT company nowadays, can prove that the testing framework is precarious. With this perspective in mind, many studies have been conducting architecture implementations for Testing as a Service (TaaS).

Those prototypes focus on concepts of static testing services fetched by internet browsers, grid-based unit testing for complex systems, parallel cloud-based automatic testings and systems that test large and complex systems with schedulers and monitoring. Furthermore some other techniques include regression testing services based on ORTS, a framework for software development and testing integration tools that facilitate development and testing processes and the Roast tool for Java API testing [5, 6].

A framework for TaaS comprises manifold complexities and manages most of the principles the testing tools mentioned so far also deal with. One of the differences is the on-demanding testing responsibility which is associated with regression testing, integration testing and performance testing. IT companies are working their development in this way so they have a continuous integration deployment service which has its testing module.

As we can see in Figure 2 a request can be submitted for testing and all performance testing procedures will be executed automatically. This advanced performance architecture increases testing efficiency and reduces time for testing software systems using automatic test case generation and test execution. To assess the workload of this system it was used Apache JMeter to execute testing scripts as parsing correctly all HTTP request info [5].

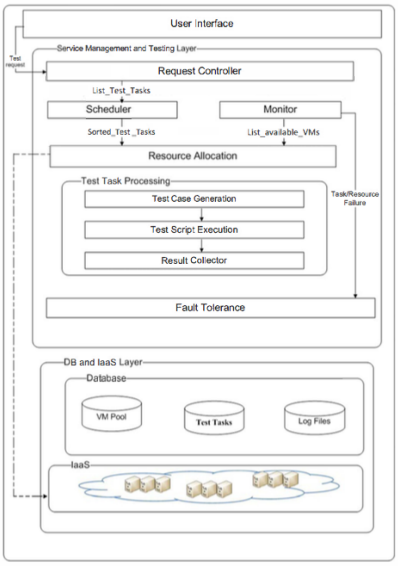


**Fig. 2:** TaaS Prototype Architecture [5]

Semi-automatic systems with a layerization of hardware and software techniques easily end up with resource bottlenecks. Usually these occur at larger number of users and it is essential to be one step ahead and identify potential faults at larger workloads. The idea is to identify at the testing time all the resources for software contentions that may become a hazard and therefore constituting a need for an information collector during the testing of multi-tier applications.

Method of Layers algorithm or also known as Layered Queuing Networks (LQN) has been addressing the performance bottlenecks for a couple of years but a predictive analyses system, based on this approach, demands at least a simulation solution of layered queuing networks [7]. Considering an architecture like this with individual queuing networks for software and hardware resources isn't enough and Duttagupta et al. [7] underlined an advanced profiling technique as one possible solution, to improve a multi-tier testing framework by estimating resource demand separately.

Most of the attempts to improve the infra-structure lead to technical barriers and that is why WAs have advantages in relation to the previous interoperability attempts such as Common Object Request Broker Architecture (CORBA). Independently of the nature of each application (finance, e-commerce, etc.) it has been the focus of research to study the performance gaps and possible mechanisms that empower enhancement. Tripathi and Abbas [8] cited some studies that presented approaches based on load testing tools in an early development cycle, mechanisms for network management and simulation network management and a framework for modeling distributed computing environments.

Every data format conversion in WAs cause serious performance bottlenecks because of the serialization/deserialization tasks on messages that are sent through the channels [8]. This is the stream encoding process every web service handles when queuing. Besides the full stack web-frameworks in focus in Section 1, there are many others open-source solutions that offer different advantages and robust micro-services to build RESTful APIs like: Beego, Gizmo, Goji, Fireball, Zeus and Volatile [9].

Companies like Huawei and Opera are using these web-frameworks in theirs development.

When doing performance testing the input parameters can be the key point of analysis to understand where the system is under-performance. When varying some parameters like the number of HTTP requests, objects size and their loading times, type of requests and concurrency we can assess how important topics as stability, scalability and throughput are behaving [10].

Patel and Gulati [10] described a correlation analysis between parameters in order to identify critical performance factors. They concluded that HTTP requests, page size and page load time are dependent parameters and they are highly correlated.

Menascé [3] also studied the performance versus load ratio criteria. He understood that with a throughput of 100 Requests per Second (RPS) and 25 concurrent requests the response time was in average about 0.25 seconds. A similar analysis showed that for 1000 virtual users the response time was about the same 0.25 seconds.

Kiran et al. [11] developed a single and distributed setup using JMeter and Unified Authentication Platform varying the number of virtual users. JMeter uses a Master and Slave servers to emulate the number of virtual users and consequently, it was possible to execute the test scripts using parameters variation. Results showed that independently of the setup, increasing the number of transactions is expected to cause a higher average throughput.

Regarding testing tools Wala and Sharma [1] presented a chart with average response times of some tools. Based on available weather web services JMeter achieved better results than other five most popularly used open source tools (Soapui Pro, Wcf Storm, Wizdl and Web Inect).

## 3 Proposed Approach

Within the API it was possible to manage and configure the HTTP layer by using an additional structure that encapsulated the native implementation with a set of new features. This study considered the usage of three different HTTP frameworks to investigate the workload of an API with multiple complex requests on an architecture with a complex data layer.

### 3.1 HTTP web frameworks

Mentioned frameworks are all open-source based tools emerged from the community to abolish technical and performance faults in the HTTP layer. The release versions used in the API were also developed in Go and were: *Revel* version v0.11.3, *Gin* version v1.0rc1 and *Echo* version v1.0.0.

With the advantage of both (API and the frameworks) being maintained in the same programming language, it was created a HTTP module for each framework that handle the configurations implementation. This module was responsible for each route specification and HTTP properties before the request was processed in the data layer. Therefore switching between frameworks was mainstream as calling the proper module.

Claimed to be the top three open-source solutions with noticed performance results (some statistics in [12] and [13]), these frameworks added some common and intuitive features like authorization layer, a new way of routing, query parameters handling, new templates with improved built-in middleware and some others [12, 13, 14].

### 3.2 API data and output

The API was connected to a Redis and PostgreSQL database. The former for fast caching of environment data preventing a higher and unnecessary overhead in each request. The latter had the data persisted in a hour basis timestamp with over one year of records for the set of possible request parameters.

API rendering module, *JSON* based, was responsible for retrieving the data (complexity-dependent on the type of time aggregation and data type request) after processing it accordingly on-demand. Listing 1.1 and 1.2 show an example of the rendered data format.

**Listing 1.1:** Level 1 data structure

```
{
  ”requestedParameter”: {
      ”timestamp”: dateTime,
      ”value”:  val
    }
}
```

Listing 1.1 renders a timestamp with a corresponding value while Listing 1.2 considers several elements for the requested parameter whereas the data field is a list of multiple elements like the one in the previous listing.

Results gathered from this study derived from a test suite with 5 different type of request parameters. Those parameters represent increasing levels of processing and *JSON* structures complexity with the heaviest one being formed by thousands of nested elements like Listing 1.1.

**Listing 1.2:** Level 2 data structure

```
{
  ”requestedParameter”:  [
    {
      ”identifier”:  val
      ”data”: [
        {
          ”timestamp”: dateTime,
          ”value”:  val
        }
      ]
    }
  ]
}
```

By the fact the API supported time aggregation (e.g: hour, day, month and year) requesting a larger period of time was the ideal pin-point case of study. Easily we could induce different behaviors into the same architecture and execute a better comprehensive load testing analysis with the number of virtual users and concurrent requests that trigger underperformance issues (see section 2).

### 3.3 Testing parameters

In section 1 we have seen some state of the art tools used for performance testing and the results achieved in order to avoid critical bottlenecks. We have also introduced the parameters analysis used in some studies as ideal conditions of testing where applications architecture usually crack: the number of virtual users at an acceptable concurrent level allows us to exploit possible throughput dropdowns and reliability issues.

The test suite used in this study considered a tuple of parameters with 1000 virtual users and 5 concurrent requests as explained in section 2. All requests were also set for one year of data and multiple time aggregations. Apache Benchmark showed to be an optimal fit permitting an easy parameterization and satisfactory report. For each request (5 in total) were collected response times from a total of 3 trials to ensure we had an average value between reasonable boundaries and that could represent effectively the production performance of the API and not just an unfortunately response glitch [15].

This study was executed in a mid-2012 machine (i5 2.7GHz Turbo Boost CPU, 8GB LPDDR3 RAM, dedicated HD GPU and 1 TB Fusion Drive HDD) with a local API instance. Databases instances were running in a different server but within the same rack. There were no significant consumption resources than the API.

## 4 Results Analysis

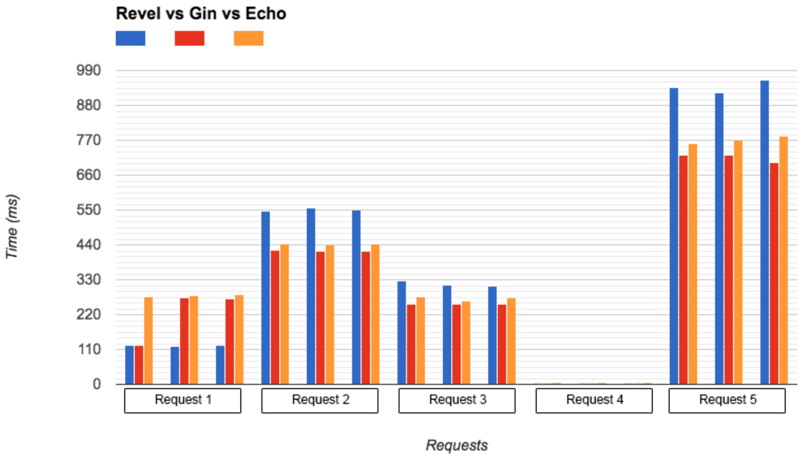


**Fig. 3:** Response time in a stack view over requests complexity

Figure 3 shows the stacked response time averages for the previous parameters and for the 3 trials for each request. Requests can be grouped in 4 types: Request 1 and 3 have similar data structures but the latter demands more complex operations in the processing layer; Request 2 doesn't charge the CPU significantly but the structure (Listing 1.2) lags the performance; Request 4 is the simplest and less demanding of all with the same data structure as shown in Listing 1.1; Response 5 is the heaviest, the more processing demanding and the one with the complex data structure.

Response time varies depending on the request, from 100ms to 990ms approximately but the best results are with Gin framework even though Echo is not that far (30ms slower). Revel prompted the slowest results with an outstanding difference in

Request 1 and 5. Request 4 was the abnormal of all, with average values of 5ms that aren't Y scale representative like the others.

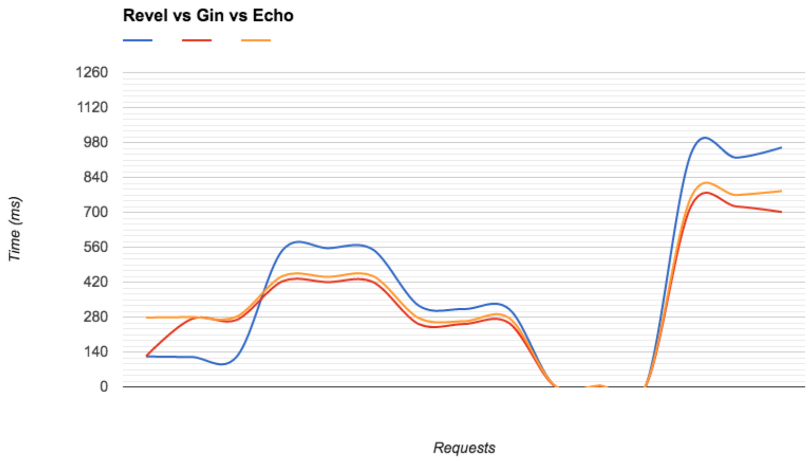


**Fig. 4:** Response time in a line view over requests complexity

Figure 4 helps perceiving the closeness between Gin and Echo in all requests and how Revel didn't get better performance except for Request 1 - this was the outlier in this study. It also provided us with a closer performance analyses showing that the API had constant performance with the three web frameworks, i.e., the follow up that lines respect across all requests match the complexity and architecture expected behavior.

## 5 Conclusions

An API with a caching system, two independent databases, a fast programming language and web-frameworks gave us impressive response time results. With a set of parameters at a concurrent level we found no stressing conditions when testing the API architecture and commuting the HTTP framework.

Our results also proved open-source community thesis that *Gin* was the fastest web-framework available. It improved our solution with notorious response time, stable HTTP layer and routing features that exploited the most of the architecture. None of the competitive frameworks provided equivalent features and better results.

## 6 Future work

With this study we see a big potential to extend this analysis over other existing HTTP frameworks that are also emerging. Additionally testing these three frameworks with different APIs available online would provide more results and possible some cutoffs in the frameworks.

Moreover re-doing this experiment with a stronger server-side architecture by running the API and the testing service in a server, would allow to analyze the impact of machine performance and network routing.

**Acknowledgments.** This work wasn't possible without the support of Movvo by giving permission to manage and collect the data from the API.

## References

1. Tanuj Wala and Aman Kumar Sharma. A comparative study of web service testing tools. 4(2):257–261, 2014. ISSN 2277 128X.
2. Igor Jugo, Dragutin Kermek, and Ana Meštrović. *Analysis and Evaluation of Web Application Performance Enhancement Techniques*, pages 40–56. Springer International Publishing, Cham, 2014. ISBN 978-3-319-08245-5. doi: 10.1007/978-3-319-08245-5_3. URL `http://dx.doi.org/10.1007/978-3-319-08245-5_3`.
3. Daniel A. Menascé. Load testing of web sites. *IEEE Internet Computing*, 6(4):70–74, July 2002. ISSN 1089-7801. doi: 10.1109/MIC.2002.1020328. URL `http://dx.doi.org/10.1109/MIC.2002.1020328`.
4. S. Sharmila and E. Ramadevi. Analysis of performance testing on web applications. *International journal of advanced research in computer and communication engineering*, 3(3), March 2014. ISSN 2278-1021.
5. A. Ali and N. Badr. Performance testing as a service for web applications. In *2015 IEEE Seventh International Conference on Intelligent Computing and Information Systems (ICICIS)*, pages 356–361, 2015. doi: 10.1109/IntelCIS.2015.7397245.
6. D. Hoffman and P. Strooper. Tools and techniques for java api testing. In *Software Engineering Conference, 2000. Proceedings. 2000 Australian*, pages 235–245, 2000. doi: 10.1109/ASWEC.2000.844580.
7. S. Duttagupta, R. Virk, and M. Nambiar. Software bottleneck analysis during performance testing. In *Computing and Communication (IEMCON), 2015 International Conference and Workshop on*, pages 1–7, 2015. doi: 10.1109/IEMCON.2015.7344508.
8. Sandesh Tripathi and S Q Abbas. Performance comparison of web services under simulated and actual hosted environments. 11(5):20–23, 2010. doi: 10.5120/1578-2110.
9. m0sth8. Awesome go, 2016. URL `https://github.com/avelino/awesome-go#actual-middlewares`.
10. C. Patel and R. Gulati. Identifying ideal values of parameters for software performance testing. In *Computing, Communication and Security (ICCCS), 2015 International Conference on*, pages 1–5, 2015. doi: 10.1109/CCCS.2015.7374125.

11. S. Kiran, A. Mohapatra, and R. Swamy. Experiences in performance testing of web applications with unified authentication platform using jmeter. In *Technology Management and Emerging Technologies (ISTMET), 2015 International Symposium on*, pages 74–78, 2015. doi: 10.1109/ISTMET.2015.7359004.
12. Gin-Gonic. Gin is a http web framework written in go (golang). it features a martini-like api with much better performance – up to 40 times faster. if you need smashing performance, get yourself some gin, 2016. URL `https://github.com/gin-gonic/gin`.
13. Labstack. Fast and unfancy http server framework for go (golang). up to 10x faster than the rest, 2016. URL `https://github.com/labstack/echo`.
14. Revel. A high productivity, full-stack web framework for the go language., 2016. URL `https://github.com/revel/revel`.
15. Tobias Fertig and Peter Braun. Model-driven testing of restful apis. In *Proceedings of the 24th International Conference on World Wide Web*, pages 1497–1502, 2743045, 2015. ACM. doi: 10.1145/2740908.2743045.

# A Survey on Mutation Testing for Database Applications

Pedro Santos

Faculty of Engineering of The University of Porto, R. Dr. Roberto Frias, Porto, Portugal,
ei11125@fe.up.pt

**Abstract.** Mutation testing is a technique to assess the quality of a test suite. This technique consists in injecting changes in the original code thus generating a modified code version called mutant. Mutation testing is an increasing interest subject and it has been applied to several areas. This paper provides an analysis and survey of mutation testing for database applications. This paper presents mutation operators and it also presents developed tools for the same purpose: assess the quality of test suites generated to testing database applications. Some analysis demonstrates that mutation testing has received further interest from researchers looking to create innovative tools and innovative approaches in order to turn this technique more applicable at an industrial level.

**Keywords:** SQL, mutation testing, database applications testing, mutation score

## 1 Introduction

Nowadays, there is a large number of database applications that make use of Structured Query Language (SQL). For this reason, to ensure the quality of these applications, there is the need to carefully ascertain the functionality of the SQL queries. The quality of an application can be ensured by applying tests to it. However, generating test cases to test database applications can be a hard process since the input information derives from several tables and each table has several rows [1].

A database application consists of two different programming languages: the host language responsible for controlling the flow of the application (e.g., Java); and the language responsible for the communication between the application and the backend database (e.g., SQL) that is embedded in the host language [2]. SQL is a declarative computer language used to query ”SQL databases” and its strength is related to the interoperability of its clauses (SELECT, FROM, WHERE, GROUP BY, HAVING and ORDER BY) and it becomes easy to formulate queries for beginners. A survey revealed that there are problems related to SQL language that make database applications appealing to use mutation testing over the generated test cases [3, 4].

Usually, software testing is considered as one phase of software development life cycle. In order to check if the developed software is meeting its requirements, the testing phase is crucial. In addition to being able to detect errors or bugs this process is also able to check if the development has taken the right path to the final solution. However, manual testing can be costly in terms of time and money. To avoid this problem, the automation of testing process has been a sought and investigated solution [5]. Therefore, the generated test cases are very important since the main goal is testing the software quality, reliability and performance.

There is a technique, called mutation testing, to assess the generated test cases. This technique generates mutants of an original program inserting syntactic modifications into it. These modifications are typically possible errors made by programmers. The generated mutants are a result of one or more than one mutations on the original program [6–9].

This paper surveys publications that take into account the use of mutation testing as technique to assess the generated test suites of database applications. Due to time limitations, the research was restricted to publications of the last five years. In addition, these publications went through a selection phase: abstract, introduction and conclusions sections were read and the publication was selected if its content was related to mutation testing for database applications. For this reason, it may be possible that some valuable publications were ignored. Thus, this survey is limited to publications related with mutation testing for database applications of the last five years.

The rest of the paper is organized as follows: Section 2 introduces concepts about mutation testing, Section 3 covers related work in the field of databases: it presents several mutation operators and developed tools. Section 4 concludes this survey.

## 2 Background

Jia and Harman [7] present a good survey and concepts analysis and we follow closely their work mainly in this section.

Mutation testing is a technique that generates and applies artificial faults into an application to improve it. Since it is difficult to generate all potential faults in a given program, mutation testing focuses only on faults that are close to the correct version of the program. The code versions where faults are applied are named mutants. Mutants can be classified into two types: first order mutants (FOMs); and higher order mutants (HOMs). FOMs are created when a mutation operator is applied just once. HOMs are created when a mutation operator is applied more than once. The injected faults try to simulate typical errors made by programmers that generate syntactic changes in the program under test (see example on Table 1). Faults are inserted within the code through a set of rules called mutation operators[1]. These rules can be increment mutator where increments are replaced by decrements and vice-versa, conditional boundary mutator where relational operators (<, <=, >, >=) are replaced by their boundary counterpart (e.g., < is replaced by <=, etc), math mutator replaces binary arithmetic operations with another arithmetic operations (e.g., + operator is replaced by - operator), etc [3, 7].

Mutation testing aims to check whether a test suite is capable to detect injected faults. A good set of test cases are those capable of detecting the inserted faults that affect the behavior of the program. If the inserted faults are detected (i.e. one of the tests fails) then the mutant is considered killed. If exist live mutants even after running the test cases, new test cases must be created to kill the mutants that are still alive. This technique generates a mutation score (MS) which is the test criterion responsible for measuring the test suite effectiveness. The goal is raise this score to 1 while the number of killed mutants increases. If this score is equal to 1 it means that the set of test cases

[1] Mutation operators are also known as mutagenic operators, mutant operators, mutation rules and mutagens [7]

**Table 1.** Example of Mutation Operation [7]

| Program *p* | Mutation *p'* |
|---|---|
| ... | ... |
| if (a>0 && b>0) | if (a>0 \|\| b>0) |
| return 1; | return 1; |
| ... | ... |

is sufficient to detect all the injected faults present in the mutants. This MS is the ratio between the number of killed mutants and the total number of mutants and it can be calculated by the following formula [8, 7, 9]:

$$MS = \frac{M_K}{M_T}, \tag{1}$$

where $M_K$ is the number of killed mutants (when different behavior between the original program and generated mutants is observed) and $M_T$ is the total number of generated mutants.

Although this technique is considered effective in assessing the quality of a set of tests, it presents some problems such as: high computational cost when executing a large amount of mutants against a set of tests; the amount of human effort that this technique involves as in the case of the human oracle problem and the equivalent mutant problem.

The human oracle problem refers to the human effort that is necessary to check the output of the original program for each test case. This problem is not particular to this technique but it is demanding to be effective and, in this case, the number of test cases can increase and therefore the oracle cost also increases.

Several cost reduction techniques have been proposed to turn this mutation testing technique into a practical testing technique. Offutt and Untch [10] divided reduction cost techniques into three types: ”do fewer”, ”do faster” and ”do smarter”. However, Jia and Harman [7] classified the reduction cost techniques into two types: ”reduction of the generated mutants” (which corresponds to ”do fewer”) and ”reduction of the execution cost” (which combines ”do faster” and ”do smarter”). They analyzed four techniques capable of reducing the number of mutants without significative loss of test effectiveness:

- Mutant Sampling, it randomly chooses a small subset of mutants from a given set;
- Mutant Clustering, it also chooses a small subset of mutants from a given set but, instead of doing it randomly, it uses clustering algorithms;
- Selective Mutation, it tries to reduce the number of mutants by reducing the number of mutation operators used;
- Higher Order Mutation, its main goal is to find higher order mutants that indicate faults. Jia and Harman [11] presented the concept of subsuming HOMs. Since HOMs are harder to kill than FOMs, it may be desirable to replace FOMs with the single HOM in order to reduce the number of mutants.

In addition to reducing the number of mutants, the computational cost can also be reduced as long as the mutant execution process is optimized. Jia and Harman [7] presented three techniques used for this optimization:

- Strong, Weak and Firm Mutation. Strong Mutation is considered as the orignal formulation of Mutation Testing, that is, for a given program *p* and a given mutant *p'*, the mutant *p'* is considered killed if it produces a different output from original program *p*. In order to optimize the Strong Mutation execution a Weak Mutation concept emerged where it is assumed that a program *p* consists of a set of components $C = \{c_1, \ldots, c_n\}$. Assuming that the mutant *p'* was generated by changing component $c_m$, mutant *p'* is killed if $c_m$ execution is different from mutant *p'*. In this particular case, instead of checking the output of the mutants after the entire program execution, mutants are checked after the execution of the mutated component. The idea of Firm Mutation is to bridge the disadvantages of both Strong Mutation and Weak Mutation. Jia and Harman claimed that, "Firm Mutation lies between the intermediate states after execution (Weak Mutation) and the final output (Strong Mutation)" [7, 12];
- Runtime Optimization Techniques. There are several techniques to do so. The Interpreter-Based Technique for instance, is used in the first generation of Mutation Testing tools. Compiler-Based Technique is another option where each mutant is compiled into an executable program and then it is executed by a set of test cases. The most recent technique is called Bytecode Translation where mutants are created from the compiled code instead of the source code. Thus, the generated mutants can be executed without being compiled;
- Advanced Platforms Support for Mutation Testing. There has been work to distribute the computational cost among different processors. Examples of this are either algorithms to improve the parallel mutation testing techniques or approaches for concurrent mutants execution under SIMD machines.

Mutation testing can be divided into two parts: Program Mutation, where the program source code is tested; and Specification Mutation, where program specification is tested.

Program mutation can be applied at both unit-level of testing (mutants represent possible mistakes made by programmers within software unit) and integration-level testing (mutants represent faults related with the connection between software units). Program mutation has been applied in several programming languages such as Java, Fortran, Ada, C, C#, SQL, etc.

Specification Mutation is about applying mutation testing at software design-level. Faults are injected on state machines or regular expressions to generate mutants called "specification mutants". A specification mutant is considered killed if its output condition is falsified. This technique can be useful either to detect missing functions in the implementation phase or to solve possible specification errors.

In addition to assessing the quality of a set of test cases, mutation testing has been applied in test data generation (the goal is to generate test data capable of killing the mutants) and has been also applied in regression testing (the idea is to apply mutation testing as a test cases prioritization technique to assist regression testing) [7].

## 3 Related Work

Since querying databases languages have begun to emerge some empirical studies have begun to be conducted to check the human performance while using different querying languages and different data models. These empirical studies revealed problems that users usually find when they are using SQL language, such as, confusing ”where”, ”group by” and ”having” clauses [1]. Taking these problems into account, applying mutation testing in order to assessing the test cases quality seems to be a fair approach.

To use this technique it is necessary to follow a set of rules. These rules are called mutation operators (as we explained in Section 2) and they can be different depending on the area where this technique is applied. Thus, we selected and present mutation operators designed specifically to be applied in database applications in subsection 3.1. Tools are also important as they seek to automate the mutation testing process (some of them generate mutants automatically) and reduce problems such as the amount of human effort and the high computational cost (as we also explained in Section 2). Therefore, we present the tools created to apply mutation testing on database applications in subsection 3.2.

### 3.1 SQL Mutation Operators

To apply mutation testing on database applications several mutation operators for SQL have been proposed. In 2005, Chan et al. [12] designed the first seven mutation operators (see Table 2) to cover insertions, replacements and deletions. All these mutation operators are detailed, explained and exemplified by Chan et al. [12].

Another set of mutation operators was proposed by Tuya et al. [1] where operators are organized into four categories: mutation of SQL clauses (SC) (see Table 3), mutation of operators in conditions and expressions (OR) (see Table 4), mutation handling NULL values (NL) since programmers and testers should avoid conditions capable of generating incorrect behaviors having null values (see Table 5) and mutation of identifiers (IR) (see Table 6).

### 3.2 Developed Tools

This section presents several tools conceived to apply mutation testing on SQL databases.

**SQLMutation Tool.** This tool[2] was developed by Tuya et al. [13] and it aims to automate the mutation process. This tool has a interface web which allows to generate mutants of SQL database queries from a web browser and it has a web service which allows to generate these mutants from other programs.

**JDAMA Tool.** Java Database Application Mutation Analyzer (JDAMA) [14] is a mutation testing tool for Java programs. This tool extends Tuya et al. approach [13] as long as it is integrated with analysis and instrumentation of the application bytecode.

[2] Tool available at *http://in2test.lsi.uniovi.es/sqlmutation*.

**Table 2.** First seven mutation operators by Chan et al. [12]

| Operator Name | Operator Description |
|---|---|
| Participation Constraint Replacement (PTCR) | This operator should mutate the must-participate entity type requirement to non-participate and vice-versa. |
| Cardinality Constraint Replacement (CDCR) | This operator changes the cardinalities of entity types relation and it also forces mutated entity to have specific values within the cardinality constraints. |
| Identifying/Weak Entity Type Replacement (IWKR) | It replaces identifying entity type by weak entity type and it replaces weak entity type by identifying entity type. |
| Attribute Replacement (ATTR) | This operator is responsible for replacing an attribute or attributes by other attribute or attributes of a compatible type. |
| Generalization/Specification Completeness Replacement (GSCR) | It replaces an expression present on a superclass by an expression present on a subclass and by the subclass negation form. |
| Generalization/Specification Disjointness Replacement (GSDR) | The main goal of this operator is to replace an expression present on a sibling entity type by an expression present on other sibling entity type as long as they share the same superclass. |
| Union Type Completeness Replacement (UTCR) | This operator changes an entity type by a subclass and/or superclass of the subclass in such a way that superclasses share the same union type constraint with the original entity type. |

**Table 3.** Mutation operators for SQL clauses (SC) by Tuya et al. [1]

| Operator Name | Operator Description |
|---|---|
| SELECT Clause (SEL) | It replaces SELECT by SELECT DISTINCT and vice-versa. |
| JOIN Clause (JOI) | This operator replaces any occurrence of JOIN keyword (INNER JOIN,LEFT OUTER JOIN, RIGHT OUTER JOIN, FULL OUTER JOIN, CROSSJOIN) by each of the others. |
| Subquery Predicates (SUB) | Subqueries can be created using predicates. There are three types of predicates, depending on the keywords type: Type I where keywords can be ALL, ANY or SOME; Type II where keywords can be IN or NOT IN; Type III where keywords can be EXISTS or NOT EXISTS. Each keyword in a predicate of any type is replaced by each of the other keywords of the same type (except for both ANY and SOME keywords as these have the same semantic meaning). In addition others replacements may occur depending on the keyword type [1]. |
| Groupings (GRU) | This operator removes GROUP BY expressions. |
| Aggregate Functions (AGR) | Aggregate functions (such as MAX, MIN, AVG, etc.) are replaced by each of the others. |
| Query Concatenation (UNI) | It replaces UNION keyword by UNION ALL keyword and vice-versa. |
| Ordering of the Result Set (ORD) | This operator replaces ASC keyword by DESC keyword to change the ordering direction and, if neither ASC and DESC keywords are present then DESC keyword is added. |

**Table 4.** Mutation operators for conditions and expressions (OR) by Tuya et al. [1]

| Operator Name | Operator Description |
|---|---|
| Relational Operator Replacement (ROR) | This operator replaces any relational operator ($=, <>, <, <=, >, >=$) by each of the others. |
| Logical Connector Operator (LCR) | It replaces AND logical operator by OR logical operator and vice-versa. |
| Unary Operator Insertion (UOI) | Each arithmetic expression related with number *n* is replaced by *-n*, *n+1* and *n-1*. |
| Absolute Value Insertion (ABS) | Each arithmetic expression related with number *n* is replaced by *ABS(n)* and *-ABS(n)*. |
| Arithmetic Operator Replacement (AOR) | Arithmetic operators ($+, -, *, /, \%$) are replaced by each of the others. |
| Between Predicate (BTW) | If a condition is "*a BETWEEN x AND y*" then it is replaced by "$a > x$ *AND* $a <= y$" and by "$a >= x$ *AND* $a < y$". |
| Like Predicate (LKE) | See rules in Tuya et al. paper [1]. |

**Table 5.** Mutation operators for handling NULL values (NL) by Tuya et al. [1]

| Operator Name | Operator Description |
|---|---|
| Null Check Predicates (NLF) | It replaces the predicate "IS NULL" by the predicate "IS NOT NULL" and vice-versa. |
| Null in Select List (NLS) | This operator transforms the items in the select list by generating mutants that will be killed when value is equal to "NULL" but it is not killed when value is not equal to "NULL". |
| Include Nulls (NLI) | It forces the condition value to true when there is a null value. |
| Other Nulls (NLO) | For each attribute *a* in *C*, the condition is replaced by "*NOT C OR a IS NULL*", by "*a IS NULL*" and "*a IS NOT NULL*". |

**Table 6.** Mutation operators for identifiers (IR) by Tuya et al. [1]

| Operator Name | Operator Description |
|---|---|
| Column Replacement (IRC) | This operator replaces each column reference by each of the other column references, parameters and constants as long as the types are compatible. |
| Constant Replacement (IRT) | It replaces each constant by each of the others constants, columns and parameters as long as the types are compatible. |
| Parameter Replacement (IRP) | It replaces each query parameter by each of the others parameters, constants and columns as long as the types are compatible. |
| Hidden Column Replacement (IRH) | The main goal of this operator is to find possible errors when many similar columns appear in the same table. |

**AGENDA Tool.** This is a set of tools for testing relational database application programs [14, 15]. It takes as input: the database schema on which the application will run; a single query that represents the application being tested; and "sample value files" that recommend values for attributes. Then, test heuristics are selected and information about the expected behaviors of the test cases is provided. With this information, this tool can populate the database, can generate inputs to the application and executes the application taking into account those inputs [15].

**MUSIC Tool.** Mutation-based SQL Injection vulnerabilities Checking (testing) is a tool that aims to automatically generate mutants for applications written in Java Server Pages (JSP) and then it aims to perform mutation analysis [16].

**MutaGen Tool.** This tool is based on SynDB framework and it handles test generation for mutation testing on database applications. It main goal is to kill mutants in database applications [17].

**iConSMutate Framework.** This framework [2] generates, automatically, test cases for database applications in both terms of coverage and mutation score. The main goal of this framework is to reduce the number of test cases avoiding test cases with low coverage and avoiding test cases with low mutation score.

## 4 Conclusions

This paper surveys and provides an analysis on mutation testing for database applications. This paper presents both mutation operators and tools designed and conceived to test databases using mutation testing technique.

After the completion of this paper we get the idea that this technique is assuming an increasing importance and researchers are trying to optimize it so that it could be applicable in industry. Several efforts have been made to reduce the cost of mutation testing. Tools created for database applications are broadening, i.e., there is a web tool, a tool for Java language, a tool for checking SQL injection vulnerabilities, etc. This demonstrates the growth of this technique in several fields of database applications.

## References

1. Tuya, J., Suárez-Cabal, M.J., de la Riva, C.: Mutating database queries. Information and Software Technology **49**(4) (apr 2007) 398–417
2. Sarkar, T., Basu, S., Wong, J.: iConSMutate: Concolic Testing of Database Applications Using Existing Database States Guided by SQL Mutants. In: 2014 11th International Conference on Information Technology: New Generations, IEEE (apr 2014) 479–484
3. McCormick, D.W., Frakes, W.B., Anguswamy, R.: A comparison of database fault detection capabilities using mutation testing. In: Proceedings of the ACM-IEEE international symposium on Empirical software engineering and measurement - ESEM '12, New York, New York, USA, ACM Press (2012) 323
4. Lu, H., Chan, H.C., Wei, K.K.: A survey on usage of sql. SIGMOD Rec. **22**(4) (December 1993) 60–65
5. Dave, M., Agrawal, R.: Search based techniques and mutation analysis in automatic test case generation: A survey. In: 2015 IEEE International Advance Computing Conference (IACC), IEEE (jun 2015) 795–799
6. Shah, S., Sudarshan, S., Kajbaje, S., Patidar, S., Gupta, B.P., Vira, D.: Generating test data for killing SQL mutants: A constraint-based approach. In: 2011 IEEE 27th International Conference on Data Engineering, IEEE (apr 2011) 1175–1186
7. Jia, Y., Harman, M.: An Analysis and Survey of the Development of Mutation Testing. IEEE Transactions on Software Engineering **37**(5) (sep 2011) 649–678
8. Strug, J.: Applying mutation testing for assessing test suites quality at model level. In: 2016 Federated Conference on Computer Science and Information Systems (FedCSIS). (Sept 2016) 1593–1596
9. Madeyski, L., Orzeszyna, W., Torkar, R., Jozala, M.: Overcoming the Equivalent Mutant Problem: A Systematic Literature Review and a Comparative Experiment of Second Order Mutation. IEEE Transactions on Software Engineering **40**(1) (jan 2014) 23–42
10. Offutt, A.J., Untch, R.H.: Mutation testing for the new century. Kluwer Academic Publishers, Norwell, MA, USA (2001) 34–44
11. Jia, Y., Harman, M.: Constructing subtle faults using higher order mutation testing. In: 2008 Eighth IEEE International Working Conference on Source Code Analysis and Manipulation. (Sept 2008) 249–258
12. Chan, W., Cheung, S., Tse, T.: Fault-Based Testing of Database Application Programs with Conceptual Data Model. In: Fifth International Conference on Quality Software (QSIC'05), IEEE 187–196
13. Tuya, J., Suarez-Cabal, M.J., de la Riva, C.: SQLMutation: A tool to generate mutants of SQL database queries. In: Second Workshop on Mutation Analysis (Mutation 2006 - ISSRE Workshops 2006), IEEE (nov 2006) 1–1
14. Zhou, C., Frankl, P.: Mutation Testing for Java Database Applications. In: 2009 International Conference on Software Testing Verification and Validation, IEEE (apr 2009) 396–405
15. Deng, Y., Frankl, P., Chays, D.: Testing database transactions with agenda. In: Proceedings of the 27th International Conference on Software Engineering. ICSE '05, New York, NY, USA, ACM (2005) 78–87

16. Shahriar, H., Zulkernine, M.: MUSIC: Mutation-based SQL Injection Vulnerability Checking. In: 2008 The Eighth International Conference on Quality Software, IEEE (aug 2008) 77–86
17. Pan, K., Wu, X., Xie, T.: Automatic test generation for mutation testing on database applications. In: 2013 8th International Workshop on Automation of Software Test (AST). (May 2013) 111–117

# Solutions for Automated Mobile Testing Cloud Service

Ricardo Lopes[1]

[1]Faculty of Engineering, University of Porto, Porto, Portugal

up201600467@fe.up.pt

**Abstract.** With the number of different mobile devices increasing every year, testing an application to make sure that it performs as expected in all devices becomes a task that can take a long time and be difficult to accomplish. Here is presented some solutions to implement an automated mobile testing process by using a mobile testing as a service approach. Taking into account the challenge of mobile testing, it is evaluated and compared some solutions that can be implemented in the development process to help automate mobile testing in multiple physical devices so that the process can be maintainable a more scalable. The comparison result shows that all solutions are viable depending on the development team requirements and that having an automated mobile testing solution implemented will decrease the time spent on the testing phase of development.

**Keywords:** Mobile testing, Mobile testing as a service, MTaaS, Mobile Testing Cloud Service, Automated Mobile Testing Service.

## 1 Introduction

As time passes the number of smartphones been used are increasing at a rapid pace. According to Satista, the number of smartphones been used by the end of the year of 2016 will reach 2.1 billion [1]. In 2011 about 10 percent of the worldwide population used a smartphone, however it is predicted that this number will increase to 36 percent by the year of 2018. Based on this trend that tends to keep on growing, companies are focusing the development of new software taking into account the mobile version. By developing apps to be used in the operating systems available for mobile devices (Android, iOS, Windows Phone, Symbian, etc.), or developing a mobile version of the companies web platform, they are increasing the number of users that the company could have. In order to show all companies and developers that they need to have in mind the mobile first approach Google as changed the way they index content, and started to index content based on a mobile first approach [2], this decision was made because they concluded that people nowadays use more mobile devices for web browsing that any other device.

Mobile development like web development have the challenge to be compatible with different operating's systems and/or devices. Like any other software it needs to be developed, tested against all the possible operating systems and/or devices where it can be used, and deployed to production. Because of the platform and device fragmentation, testing the application can take a very long time and it is not possible for most companies to have different types of devices to guarantee that the application will perform well in the different operating's systems versions, screen size, resolution and models [3] [4]. As the number of devices and OS versions grow, mobile testing becomes more demanding and maintaining the application reliability and security becomes a task very difficult to accomplish. To guarantee that this does not happen, it is necessary to have a service where we can automate the testing process on a variety of devices and/or OS.

In order to satisfy the demand of having a scalable mobile testing process comes the concept of mobile testing as a service (MTaaS). Has been clearly stated by [5], MTaaS provides on demand testing services for mobile applications to support software validation and quality engineering processes by leveraging a cloud-based scalable mobile testing environment to assure pre-defined given quality of service requirements and service level agreements (SLAs). MTaaS helps reduce the high costs in current mobile testing practice and environments, gives more testing support and tools for mobile scalability test, and reduces the high mobile testing complexity and harness due to high diversity in mobile devices, platforms, browsers and environments.

In order to facilitate and improve the testing phase of mobile applications, this paper will present the concepts, advantages and possible solutions of using MTaaS to improve this phase of development, making the testing process more reliable, scalable and less time consuming so that the software development process can be one step closer to continuous delivery [6]. To help implement the process, four solutions will be analyzed and compared. The results presented will assist in the decision of which solution is the ideal for certain scenarios.

The rest of the paper is organized as follows. In the second section it is discussed the mobile application testing and the MTaaS model. In the third section it is presented the available solutions for mobile applications testing using cloud-based services. In the fourth section the results of a comparison between the solutions is presented. The fifth section provides an overview of research studies made in the mobile testing service area. Finally, in section six it is presented the conclusions and future works.

## 2 Mobile Testing as a Service

A new service model called Testing as a Service (TaaS) has gained a lot of popularity over the last few years because of its scalability. The idea behind this concept is to have a cloud based service to provide testing capabilities to users [9]. This model offers software testing as an economical, effortlessly accessible web service and entirely automated testing in the cloud. Mobile Testing as a Service combines all the aspects of the TaaS adding the complexity of testing the software on various mobile devices and operating systems.

With the rapidly growing of different types of mobile devices and their features it becomes very difficult to test the application in a large number of smartphones. Mobile Testing as a Service is a model that makes that task achievable, by offering a service where a user can send their application and tests that need to be run, and the service installs the applications in multiple physical devices, run the tests, and then returns to the user a report containing the result of the tests in every device.

In Figure 1 the workflow of the whole process with a CI system is exemplified. First a developer checks in their new version of the application to a code repository. Then the CI system creates a new build of the application and uploads it to the test cloud. The test cloud deploys the application to multiple devices with different characteristics (size, features, OS versions, etc.) and runs the tests on every device. When all the tests are completed a report containing the results of the tests in all devices is returned to the end user.

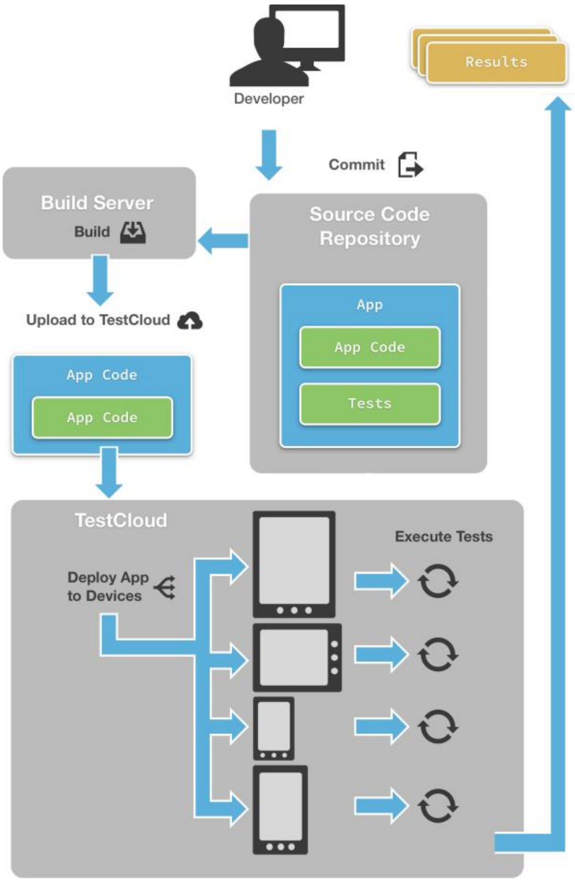


**Figure 1.** Mobile Testing as a Service with Continuous Integration

According to Mona Erfani [6] and C.Mano [10], MTaaS is important due to the following reasons: cost reduction by leveraging cloud computing resources and mobile device distribution, recurrent variations and upgrades of mobile platforms and technologies, high mobile test complexity, mobile scalability of software as a service (SaaS) and applications, on-demand testing service, mobile usability and internationalization, and testing multi-tenancy of mobile SaaS.

## 3 Mobile Testing as a Service Solutions

Most of cloud services for mobile testing provide a group of devices where an application can be deployed to and tested. These devices are hosted by the cloud services, and remote access to those devices is given in order to perform all the testing needed to approve the application. These services offer the possibility of interaction with the devices using an interactive interface, and automation of running tests in multiple devices at the same time.

### 3.1 Xamarin Test Cloud

*Xamarin Test Cloud* (www.xamarin.com/test-cloud) provides a service where an Android or iOS application can be uploaded to be tested in more than 2400 different devices.

Written tests using the integrated template can be used to automate all tests to be run in all devices. The devices contain different versions of the Android and iOS operating system and any device feature can be used during the test (GPS, camera, rotation of the device). The service can be integrated with different continuous integrated (CI) systems (Team Foundation Server, Jenkins, and TeamCity) and also generates reports about the performance of the application (memory, CPU, duration, etc.).

The Xamarin Test Cloud service is consisted of four essential parts:

- A framework called Calabash so that developers can write tests with Cucumber (www.cucumber.io) and Ruby (www.ruby-lang.org). The tests are well suited to Behavior Driven Development (BDD), which helps focus development on the delivery of prioritized, verifiable business value by providing a common vocabulary that reduces the boundary between business and technology. BDD relies on the use of a very precise vocabulary to diminish miscommunication and to guarantee that everyone are not only on the same page but using the same words [17].
- A framework called Xamarin.UITest that lets developers write tests with the C# language using the testing library NUnit (www.nunit.org). This is ideal for teams that are familiarized with written test with NUnit and/or are already doing their mobile developments using Xamarin (www.xamarin.com).
- A test cloud service that has thousands of physical mobile devices of different hardware, sizes, features, and OS. The application that needs to be tested is uploaded to the service, which will install it on all the devices, run the tests, and returns and informative report to the user with the results.

- A tool (Xamarin Test Recorder) to help users that are not familiarized with written tests or using APIs. A user starts the tool, connects to a device and performs interactions in order to create a test workflow. The tool will capture all interactions and then create a scenario written in C# so that it can be included in the group of tests to be run.

This service provides all the necessary components that allows teams to create automated tests to be run on multiple devices and verify if the behavior was accurate. In Figure 2 is presented the test cloud service workflow.

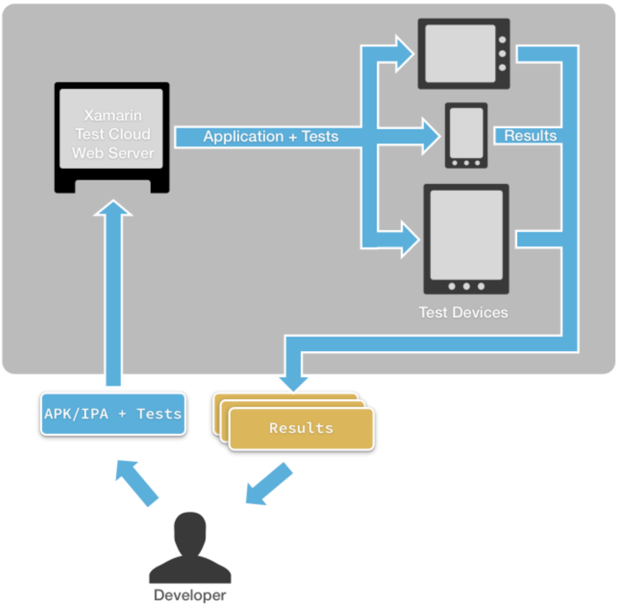


**Figure 2.** Test Cloud Workflow

The whole process is very well implemented and can handle the more common testing scenarios, but at this moment, the service can't run tests that need the following requirements:

- Run tests for more than 180 minutes
- Testing with Bluetooth connected
- Network connection that isn't Wi-Fi
- Throttling Wi-Fi
- Different battery setups
- Camera usage
- Starting an application automatically rotated
- VPN into corporate network instead of opening up ports to the firewall
- Integration with other applications installed on the device

### 3.2 AWS Device Farm

*AWS Device Farm* (www.aws.amazon.com/device-farm/) offers a service to test an application with different operating systems (Android, iOS) and also web apps on more than 350 different devices. The testing process can be automated by selecting the tests that will be executed using a continuous integrating environment like Android Studio and Jenkins. A test environment can be configured to simulate real-world conditions like language, location and network connection. Reports with the application performance and other metrics are given after the tests have been run to help identify problems with the application.

This service helps users to interact and test a mobile application with different mobile devices and OS that are hosted in Amazon Web Services (AWS). A user can use the service by automating the applications tests using a testing framework or by remote accessing a device to perform the necessary interactions to validate the status of the application. The tests are run in parallel and as the tests complete a report containing screenshots, performance data, low level logs and high level results. The service supports testing of native and hybrid Android (with remote access for interactive testing), iOS, and Fire OS apps, including those created with Xamarin, PhoneGap (www.phonegap.com), Titanium (www.appcelerator.org), Unity (www.unity3d.com) and other frameworks.

Remote access to test an application allows user to interact with the device in real time, which can be useful in some situations. During the remote access sessions logs, action detail and a video of the session is produced at the end.

The main limitations of the service are listed below:

- Limited number of Android and Fire OS devices are supported for remote access
- Maximum app file size is 4 GB
- Maximum 5 devices to run tests in parallel
- 60 minute remote session limit

### 3.3 Appium

*Appium* (www.appium.io/) is an open-source tool for automating applications on iOS and Android operating systems. Tests can be created using different languages and then sent to the server to be executed. It can be installed on a private server and it creates a device hosting environment that can be integrated with multiple CI systems to automate the run of tests on multiple devices.

This tool allows the writing of tests against multiple platforms (iOS, Android, Windows), using the same API, which enables code reuse. The service uses vendor-provided automation frameworks so that there is no need to compile that application in any specific way. It has a client-server architecture so that a testing client can be created in any language, it only needs to follow a specific protocol and send the appropriate HTTP requests to the server.

Appium can be entitled a webserver that exposes a REST API. It accepts connections from a client, listens for commands, performs those commands on a mobile device, and responds with an HTTP response representing the result of the command execution. Existing a client/server architecture opens the possibility to write the test code in any language that has a HTTP client API, the server can be on a machine while the tests are running on a different machine, write test code and rely on a cloud service to receive and interpret the commands. The automation is always performed in the context of a session. Clients initiate a session with a server in ways specific to each library, sending a POST /session request to the server, with a JSON object called the 'desired capabilities' object. At this point the server will start up the automation session and respond with a session ID which is used for sending further commands. Desired capabilities are a set of keys and values sent to the server to tell the server what kind of automation session is intended in starting up. The server is written in Node.js (www.nodejs.org). There are client libraries (in Java, Ruby, Python, PHP, JavaScript, and C#) which support Appium's extensions to the WebDriver protocol. There are graphical user interfaces (GUI) wrappers around the server that can be downloaded. These come bundled with everything required to run the server, so there is no need to have some apprehension about Node. They also come with an Inspector, which enables you to check out the hierarchy of your app. This can come in handy when writing tests.

### 3.4 Calabash

*Calabash* (calaba.sh/) is an open-source tool for automating acceptance testing for mobile applications. It provides a framework for Android and iOS in order to create the tests and run them on the devices. The tool can be used in different CI systems and is maintained and used by Xamarin on their MTaaS solution.

It consists of libraries that enable test code to interact programmatically with native and hybrid apps. The interaction consists of a number of end-user actions. Each action can touches, gestures, swipe, rotate, assertions, and screenshots.

Calabash has support for Cucumber. Cucumber lets you express the behavior of your app using natural language that can be understood by business experts and non-technical quality assurance (QA) team. Although Cucumber is the main focus, Calabash can be used to write automated functional and acceptance tests using any Ruby-based test framework.

## 4 Results

Table 1 summarizes the MTaaS solutions mentioned above and a comparison based on their main features.

**Table 1.** List of MTaaS Solutions

| | Xamarin Test Cloud | AWS Device Farm | Appium | Calabash |
|---|---|---|---|---|
| Open Source | | | X | X |

| | | | | |
|---|---|---|---|---|
| Commercial | X | X | | |
| Android | X | X | X | X |
| iOS | X | X | X | X |
| Other OS | X | X | X | |
| Different OS versions | X | X | X | X |
| Number of devices | >2400 | >350 | | |
| Max testing time in minutes | 180 | 60 | None | None |
| Device Interaction | X | X | X | X |
| Testing Results | X | X | X | X |
| Performance Results | X | X | | |
| CI support | X | X | X | X |

All the services mentioned in table 1 can be applied to a mobile developing process. Each of them have their advantages and disadvantages, but all of them can achieve the goal of making the mobile application development scalable, maintainable and more reliable.

The commercial solutions have the advantage that starting to use the service will be more rapidly since all the infrastructure is already in place and only minor account configurations need to be done. On the downside there are some limits in the number of devices available, and the tests need to be carefully selected to take the least amount of time possible.

The open source solutions have the downside that all the configuration, installation and integration has to be done by the development team, and the mobile devices need to be acquired, but all the system is controlled by the team, so the tests can be run at any time and take the time necessary to complete, also any custom configurations or implementations are more easily implemented.

Automating mobile testing improves the productivity of applications being developed, but manual testing is still needed and there is room for manual and automated testing [7] [8]. Automating test cases helps insure that those tests are run when new versions of the applications are developed in a shorter time span. When all this process is automated and run on a CI system, developers and quality assurances can have more time to identify new test cases to be implemented and do more exploratory tests trying to identify more possible problems. This will help have a scalable and maintainable process as more and more tests are automated.

## 5 Related Work

Several works are being done to help provide better solutions for TaaS for all OS.

Shenbin Zhang [12] provides mobile application functional test solution based on TaaS platform. The solution focus on Android native apps and provides an integrated app test environment. Using an automatically generated test script based on functional traversal, the TaaS user can customize the test environment by configuring the devices in the cloud. After the project is submitted the TaaS can execute the distributed test automatically.

Chuanqi Tao [13] presents a setup for a Mobile Infrastructure as a Service (MIaaS) for MTaaS. It is presented a prototype of MIaaS for MTaaS with support for resource provisioning, monitoring, billing services, and a consistent mobile test infrastructure for different mobile testing service tools for the future.

Jun-fei Huang [14] presents a Remote Mobile Test System, designed to solve the problems associated with maintaining real mobile phone and inefficient compatibility testing. The system is essentially a mobile testing cloud, and end users can request devices on demand to test their software with automated testing features.

C.Mano Prathibhan [15] proposes a cloud based mobile application framework that can be used to perform various type of testing in any given Android mobile application. The framework has an automated testing tool that integrated with the testing framework provides Android Testing as a Service (ATaaS) for users.

Tao Zhang [16] provides a statistical and step-by-step approach for engineers to obtain an optimal test sequence for diverse mobile test devices, platforms, and environments. It discusses a tree model to support compatibility test modeling and analysis, a statics method applied to cluster mobile devices with similar configurations and appliances to avoid redundant testing, an optimization strategy to rank mobile devices by their market share and a compatibility testing service.

The list of studies presented are a combination off areas that integrate with mobile testing as a service.

## 6 Conclusions

Guaranteeing the quality of mobile applications is difficult because of the number of variables that need to be considered. Multiple hardware and software versions, network interfaces, existence of different type of mobile applications (hybrid, web), diversity of sensors, cameras, battery, etc. Because of all of these factors that need to be taken into account, testing mobile applications needs to be a thorough process and of very high importance for all parties involved in the development, testing and consumer phases.

For these reasons, extensively testing is migrating to the cloud so that the testing process is done in less time and to decrease development expenses, that is very important for several companies that don't have the budget to have a bigger development team.

In the paper it was described the Mobile Testing as a Service workflow and presented its main advantages of having a cloud based testing system. It was presented some solutions to start implementing this process, each solution has its advantages and disadvantages and it is up to the person responsible for the development process to choose which solution is the one that fits the way of work of its development team in order to get the most out of this service.

Mobile testing clouds, on-demand mobile testing services infrastructures, and mobile test automation solutions are needed to meet the demands on mobile testing.

## References

1. Gao, Jerry. "Mobile Testing-As-A-Service (MTaaS) - Infrastructures, Issues, Solutions and Needs." *International Symposium on High-Assurance Systems Engineering*. Miami, 2014.
2. Gao, Jerry, K. Manjula and P. Roopa. "A cloud-based TaaS infrastructure with tools for SaaS validation, performance and scalability evaluation." *4th International Conference on Cloud Computing Technology and Science*. Taipei, 2012.
3. George Candea, Stefan Bucur, Cristian Zamfir. "Automated Software Testing as a Service." *ACM symposium on Cloud computing*. Indianapolis, 2010.
4. Gligoric, Milos. "An Empirical Evaluation and Comparison of Manual and Automated Test Selection." *Automated Software Engineering*. Vasteras, 2014.
5. Huang, Jun-fei and Yun-zhan Gong. "Remote mobile test system: a mobile phone cloud for application testing." *4th International Conference on Cloud Computing Technology and Science*. Taipei, 2012.
6. Joorabchi, Mona Erfani. "Real Challenges in Mobile App Development." *International Symposium on Empirical Software Engineering and Measurement*. Baltimore, 2013.
7. Lehtonen, Timo. "Defining Metrics for Continuous Delivery and Deployment Pipeline." *Symposium on Programming Languages and Software Tools*. Tampere, 2015.
8. *Number of smartphone users worldwide from 2014 to 2020 (in billions)*. 2016. <https://www.statista.com/statistics/330695/number-of-smartphone-users-worldwide/>.
9. Phan, Doantam. *Mobile-first Indexing*. Friday November 2016. <https://webmasters.googleblog.com/2016/11/mobile-first-indexing.html>.
10. Prathibhan, C.Mano, A. Malini and N. Venkatesh. "An automated testing framework for testing Android mobile applications in the cloud." *International Conference on Advanced Communication Control and Computing Technologies*. Panipat, 2014.
11. Sharma, R. M. "Quantitative Analysis of Automation and Manual Testing." *International Journal of Engineering and Innovative Technology* 4.1 (2014): 6.
12. Shevchik, Leigh. *Top 3 Challenges Facing Mobile Development*. 15 November 2012. <https://blog.newrelic.com/2012/11/15/top-3-challenges-facing-mobile-development/>.
13. Solís, C. and X. Wang. "A study of the characteristics of behaviour driven development." *Software Engineering and Advanced Applications*. Oulu, 2011.
14. Tao, Chuanqi, Jerry Gao and Bixin Li. "Cloud-Based Infrastructure for Mobile Testing as a Service." *Third International Conference on Advanced Cloud and Big Data*. Las Vegas, 2015.
15. Zhang, Shenbin and Bingfeng Pi. "Mobile Functional Test on TaaS Environment." *Symposium on Service-Oriented System Engineering*. Oxford, 2015.

16. Zhang, Tao, et al. "Compatibility Testing Service for Mobile Applications." *Symposium on Service-Oriented System Engineering*. San Francisco Bay, 2015.

# Common Threats in Security Testing with Scrum Framework: A Systematic Literature Review

Thiago Baldês[1]

[1] Department of Informatics Engineering
Faculty of Engineering of University of Porto, Porto, Portugal
up201600468@fe.up.pt

**Abstract.** The Agile methodology has a remarkable use in software engineering. Especially Scrum framework has a huge adoption in market by the enough recognition as efficient development process in by delivering software fast even under the time constrains. However, like other agile methods including, Feature Driven Development (FDD) and Extreme Programing (XP) has also been criticized because of unavailability of effective security elements in their practices. Moreover, security of web applications is still a major issue of increasing importance due to the growth of the Internet. In order to have a deeper look into the matter and understand more about the reality, we conducted a systematic literature review (SLR) and studied the literature and software solutions for the last 10 years for analyze what are the mains issues in agile security testing. The result of this paper shows the important aspects to be considered in terms of the organizational point, the project, the team and the process during application of the Scrum.

**Keywords:** Agile Methodology, Scrum, Security Testing, Software Engineering

## 1 Introduction

Agile methodologies have had an important impact on software development practices in recent years. A significant amount of positive feedback has been noted from the organizations that practice agile methods in software engineering. Their statements [1, 2] suggest that agile methods help during the software development process by emphasizing rapid development. This, along with an ability to quickly respond to changes in requirements, leads to a high degree of customer satisfaction. Agile methods are more flexible and help to reduce iterations. However, they need to follow several rules related to the agile manifesto, including those concerning less documentation and team member interactions, which provide appropriate communication with customers and other users.

On the other hand, the security of web applications is still a major issue of increasing importance due to the growth of the Internet and the vast amount of available applications and services. Ensuring security of applications requires thoroughly testing at all stages of the development process, and therefore, several testing methods have been investigated in order to optimize the security testing process. Nevertheless, agile methodologies do not suggest or include security elements in their models. In general, the exclusion of security elements from the agile development process creates vulnerable software.

Although the Scrum framework is the most popular agile methodology in world organization, it does not provide guidelines for dealing with the security aspects of software [3]. The original Scrum model does not include software security planning from the start [4]. This means that security practices are not expected to the Scrum team. For instance, security reviews are supposed to be completely objectively performed by the use of automated tools to compensate the lack of effective security testing during the development.

In this paper we present a Systematic Literature Review (SLR) realized toward identifying common problems, threats with Scrum framework. In this way, we can correlate the practices of this approach with vulnerabilities of the web application in order to analyze the security testing perspective. This paper is organized as follows. Section 2 gives a brief description of Scrum methodology and the agile testing. Section 3 describes the method used for the systematic literature review (SLR). Section 4 exposes the results of the SLR. Section 5 presents the conclusions and future work.

## 2 Main Concepts

Bringing the security approach in the context of agile software development, this paper describe a Scrum Framework as the agile methodology and security testing in that environment as main concepts.

### 2.1 Scrum

Scrum is the most popular agile methodology in the software industry and using those Scrum practices, several companies have improved their quality and productivity [5]. The core of the Scrum framework is that customer requirements can be changed during development and software is developed iteratively. Iterations are called sprints and every sprint starts with a sprint planning meeting where the customer reviews and prioritizes requirements.

The prioritized requirements recorded as user stories that can be delivered by the development team are selected, agreed and transferred to sprint backlog that is used to manage the requirements during the sprint. Then the development team works together to develop software features, satisfy customer requirements and deliver shippable software by the end of each sprint. This shippable software increment is presented in the sprint review meeting where the customer can think about further requirements. The most frequent meetings are daily stand-up meetings where development team members discuss what they have done since the last meeting, what they will do in the coming days and whether they are facing any impediments.

Three roles are defined in the Scrum framework. The Product Owner represents customers and is responsible for defining and prioritizing software requirements and records them in the product backlog. The Development Team is responsible for delivering the potentially shippable software by the end of each sprint. The Scrum Master facilitates Scrum meetings and ensures the development team can work as efficiently as possible [6]. In Fig. 1 we can observe the Scrum process, showing the most basic elements that characterize this methodology.

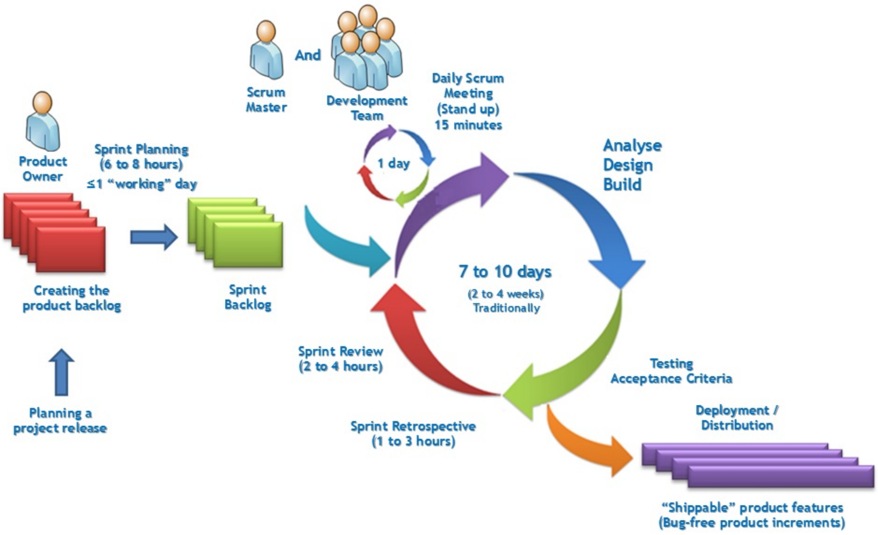

Scrum Master
And
Development Team
Daily Scrum Meeting (Stand up) 15 minutes
1 day
Product Owner
Sprint Planning (6 to 8 hours) ≤1 "working" day
Analyse Design Build
Creating the product backlog
Sprint Backlog
7 to 10 days
(2 to 4 weeks) Traditionally
Sprint Review (2 to 4 hours)
Testing Acceptance Criteria
Planning a project release
Sprint Retrospective (1 to 3 hours)
Deployment / Distribution
"Shippable" product features (Bug-free product increments)

**Fig. 1.** Scrum process[1].

### 2.2 Agile Security Testing

For competitive reason such as short delivery, flexibility, cost control, transparency, agile methods are widely used for web and network application development [7]. Regardless of security risks, the agile methods have few features specifically addressing security issues [8].

The Agile Security Testing methodology, suggested by Tappenden et al. [9], consists of three main steps. Step 1, the modeling of security requirements, is

[1] http://www.scrumguides.org/scrum-guide.html#definition

executed by creating abuser stories [10] and/or misuse cases [7, 8] in order to elicit security requirements. These are then used as reference points when testing for security in order to verify or falsify a given security requirement. Step 2 a highly testable architecture, is achieved by adding a test layer on top of each of the three layers that Web applications typically consist of, for example, presentation layer, business service layer and data service layer. Step 3, running automated security tests, which is necessary in order to fully benefit from Agile Security Testing.

Researchers suggest that there is a clash or disparity between established security engineering processes and methodologies proposed by the agile manifesto. This implies that agile methods work with short development increments that adapt easily to change, whereas existing security engineering processes rely on a waterfall development setting to reduce defects through heavy and inflexible processes. However, the growing trend towards the use of agile techniques for building software and the immense increase in security breaches over the past few years imply that it is essential to integrate the existing engineering processes with agile process.

## 3 Method

In this paper was used a systematic literature review as research method. According to Kitchenham [11], this method is for identifying, evaluating and finally interpreting literature relevant for research questions in a certain topic area. It is one of the most important methods in Evidence-Based Software Engineering (EBSE) [12]. The systematic review reported in this article has the objective of summarizing the main studies about the security testing with agile methodology focus on Scrum in software engineering. The following sections will provide an overview of those approaches, regarding the relevant concepts found for this literature review.

### 3.1 Research Objectives and Questions

The purpose of this systematic literature review is to analyze scientific publications in order to identify security perspective in organization during production of their products by Scrum methodology. It means, analyze security testing in the agile practices considering the steps of the process. The research questions that guided the systematic review process, were:
Q1: What are the main issues of security testing in Scrum?
Q2: What is the impact on agile development especially in Scrum?

### 3.2 Search Strategy

The strategy used to construct search terms is as follows:

*a)* Derive the major terms from the research questions by identifying the population, intervention, context, and outcome;
*b)* Building search strings using AND and OR to include synonyms and abbreviations and to join the main terms derived from the research questions;
*c)* Dividing in 4 phases the search with different context according to the tables 1, 2, 3 and 4;
*d)* The main fields during the phases are: Security Testing - ST, Software Development - SD, Agile - A and Scrum – S;
*e)* The contexts are: TAK - Title, Abstract and Keywords; K - Keywords only.

**Table 1.** Search String Phase 1

| Steps | Field 1 | Context 1 | Field 2 | Context 2 | Output |
|---|---|---|---|---|---|
| 1 | ST | TAK | SD | TAK | 1594 |
| 2 | ST | TAK | SD | K | 428 |
| 3 | ST | K | SD | TAK | 540 |
| 4 | ST | K | SD | K | 222 |

**Table 2.** Search String Phase 2

| Steps | Field 1 | Context 1 | Field 2 | Context 2 | Output |
|---|---|---|---|---|---|
| 1 | ST | TAK | A | TAK | 122 |
| 2 | ST | TAK | A | K | 46 |
| 3 | ST | K | A | TAK | 25 |
| 4 | ST | K | A | K | 18 |

**Table 3.** Search String Phase 3

| Steps | Field 1 | Context 1 | Field 2 | Context 1 | Field 3 | Context 3 | Output |
|---|---|---|---|---|---|---|---|
| 1 | ST | TAK | SD | TAK | A | TAK | 66 |
| 2 | ST | TAK | SD | K | A | K | 34 |
| 3 | ST | K | SD | K | A | K | 26 |

**Table 4.** Search String Phase 4

| Steps | Field 1 | Context 1 | Field 2 | Context 2 | Output |
|---|---|---|---|---|---|
| 1 | ST | TAK | S | TAK | 22 |
| 2 | ST | TAK | S | K | 18 |
| 3 | ST | K | S | TAK | 16 |

| | | | | | |
|---|---|---|---|---|---|
| 4 | ST | K | S | K | 13 |

### 3.3 Study Selection

In this work, we decided to consider a large diversity of studies related to the research topic, including recent papers journals, proceedings of conferences and congresses in technology area. The search selection is limited to the following database sources:

- Scopus
- ACM Digital Library
- Springer Link
- IEEE Xplore

The relevance of the studies for this work was evaluated through the use of the inclusion criteria and the study's quality assessment criteria. Once potentially relevant studies were obtained from the searches, they were deeper analyzed and the relevant studies were included and those with little relevance were discarded.

## 4 Result

In our findings, were identified common problems that threat the security testing in a web project with agile methods as development paradigm. Moreover, was noted some issues especially to the Scrum approach as well. This section presents the systematic literature review results, organized according to the research questions and for it, we put the answers in four categories: organizational, the product, the team and the process.

*Organizational*

In any company, the success of the product depends on the many factors. The papers [5, 13–16], agree that organizational culture has the great importance for it. In this case, the support from the stakeholders, transparency, to walk toward of the mission contributes for positive aspect. The Agile methodologies must be used within an agile culture that is characterized by a broad support for the negotiations, a capacity for change, the collaboration and the continuous exchange of experiences and knowledge.

*Product*

This category was related to vulnerabilities against the companies' product. Most of the agile security testing facing threats web application. In many papers were identified security problems been common nowadays. According to the paper [5, 17–20], the OWASP, the top 10 vulnerabilities in web application are:

i. SQL Injection
ii. Broken Authentication and Session Management
iii. Cross-Site-Scripting(XSS)
iv. Insecure Direct Object Reference
v. Secure Misconfiguration
vi. Sensitive Data Exposure
vii. Missing Function Level Access Control
viii. Cross-Site Request Forgery (CSRF)
ix. Using Components with Known Vulnerabilities
x. Unvalidated Redirects and Forwards

Most of the time, in agile development is adopted the use of automated testing tools in advanced perspective for those threats, in order to ensure a high coverage of project tests. During the development step, the security is checked on business logic, looking for misuse scenarios for features.

*Team*

Looking deeply, one of the key points that we found was that the agile approach increased the security awareness in the development team, not only in tester team. Sometimes developers and architects are less prone to worry about security for a number of reasons. Raising security awareness in the development team was hence a considerable benefit. The development team became accustomed to asking the question "how can this functionality be misused?"

Studies as [5, 13, 21, 22], are taking a look to the problem as a lack of team training, communications through the roles involved, size of the team, teams not aligned, the skills of the teams, rotation of members of the team, lack of experience with the agile methods, lack of effective communication and misunderstandings, security awareness and lack of comprehension of the Agile values.

*Process*

For this category, we found the papers [5, 8, 13, 14, 20, 23]which confront the security agile testing in Scrum frameworks in some aspects:

- Software security inspections: Where security testers manually review the surface of the software solution, understanding and checking its client-server protocols, resources, features, and functional security controls;
- Security checks on business logic: Making use of implemented functionalities looking for misuse scenarios in front of, such as impersonation, elevation of privileges, privacy breaches.
- Automated web vulnerability analysis: Executing security scanners against the targeted web application dynamically when the software is running. Automated source code analysis.
- Manual web vulnerability analysis: Performing manual attacks against the targeted web application with the software running, just using a web browser or a web proxy as tools. Manual code reviews of top-critical parts of the application.
- Third-party dependencies analysis, looking for published vulnerabilities that could affect the application security.
- Restrict information: Unauthorized access Log analysis, verifying that accurate, proportional, and appropriate information is properly registered in the software traces, and that no sensitive information is being logged by mistake.

In the Scrum sprint approach, the users stories are what drive the development, but there is no focus on security for it in this step. The studies 13], [14], [20], revealed relevant points in this agile methodology:

- Security tests after the first sprint: In addition to the usual activities, which are common at the beginning of the testing phase, the objective of this stage is to configure the security testing environment needed for the rest of the stages, it included the configuration of the automated tools for dynamic analysis and source code analysis.
- Security tests for every subsequent sprint: Every automated tool configured after the first sprint is executed at least once per sprint, looking for software implementation errors that could lead to vulnerabilities related to most common attacks as mentioned by OWASP in project category.
- Security tests on every release: Before the delivery of a new release, in addition to the "on every sprint" tasks, a set of security tests should be taken. Manually verified executed by the security Team. The automated tools also can be used in a dynamic context involving regression and integration testing

Finally, the table 5 shows the 20 papers selected organized by the categories of potential problems found in our study. This means, for each paper analyzed, we assigned according to the type of the threat.

**Table 5.** Papers Selected

| | **Categories** | | | |
|---|---|---|---|---|
| **Reference** | ***Organizational*** | ***Product*** | ***Team*** | ***Process*** |
| [17] | | x | | |
| [18] | | x | | |
| [15] | x | | | x |
| [8] | | x | | x |
| [23] | | | | x |
| [24] | | x | | |
| [25] | x | x | x | x |
| [13] | x | | x | x |
| [14] | x | | | x |
| [26] | | x | | |
| [27] | x | | | x |
| [28] | x | | x | x |
| [21] | | | x | |
| [20] | | x | | x |
| [19] | | x | | |
| [22] | | | x | |
| [29] | | x | | x |
| [16] | x | | | |
| [5] | x | x | x | x |
| [30] | x | | | x |

## 5 Conclusion and Future Work

This study presents a systematic literature review on security testing in Scrum. Were pointed 20 papers with relevance on the research questions and the study criteria.

Base on our finding of this research, we regarded the view through the organizational, the project, the team and the process perspective involved in our work. After the review of the papers, we could identify common threats and put in categories. The use of the automated tools has a important role during the agile process been used to test automatically analysis of the code, browsing testing verifying the vulnerabilities and to measure the performance of web application with satisfactory coverage. On the other hand, it does not imply on manual tests that also can be combined in order to quality assurance.

Although the Scrum is the most used agile methodology in the companies, the agile security testing presented considerable issues that are potentially discussed from the papers analyzed. This work organized by categories the result of this analysis focus on evidences of systematic literature review, it means, security testing activities synchronizing with agile methodology of the Scrum. Is extremely important the security awareness in the whole cycle of the methodology and not be restrict only by security team.

As future work, we can enlarge this systematic literature review for other scientific data sources such as Web of Science, Google Scholar and other relevant area repository, in order to bring other evidences. Also to apply this security approach used in Scrum in others agile methods like Extreme Programming XP and Feature Driven Development FDD.

## References

1. Azim Sharifloo A, Saffarian AS, Shams F (2008) Two architectural practices for extreme programming. Commun Comput Inf Sci 6 CCIS:964–967. doi: 10.1007/978-3-540-89985-3_144
2. Breivold HP, Sundmark D, Wallin P, Larsson S (2010) What does research say about agile and architecture? Proc - 5th Int Conf Softw Eng Adv ICSEA 2010 32–37. doi: 10.1109/ICSEA.2010.12
3. Ghani I, Azham Z, Jeong SR (2014) Integrating software security into agile-Scrum method. KSII Trans Internet Inf Syst. doi: 10.3837/tiis.2014.02.0019
4. McGraw G, Allen JH, Mead N, et al (2013) Software Security Engineering: A Guide for Project Managers. Softw Eng Inst. doi: 10.1017/CBO9781139083683
5. Lopez-Martinez J, Juarez-Ramirez R, Huertas C, et al (2016) Problems in the adoption of agile-scrum methodologies: A systematic literature review. Proc - 2016 4th Int Conf Softw Eng Res Innov CONISOFT 2016 141–148. doi: 10.1109/CONISOFT.2016.30
6. Schwaber K, Sutherland J (2013) The Scrum Guide. ScrumOrg and ScrumInc 17. doi: 10.1053/j.jrn.2009.08.012
7. Sindre G, Opdahl AL (2005) Eliciting security requirements with misuse cases. Requir Eng 10:34–44. doi: 10.1007/s00766-004-0194-4
8. Erdogan G, Meland PH, Mathieson D (2010) Security Testing in Agile Web Application Development-A Case Study Using the EAST Methodology. Agil Process Softw Eng Extrem Program 14–27. doi: 10.1007/978-3-642-13054-0_2
9. Tappenden A, Beatty P, Miller J, et al (2005) Agile security testing of web-based systems via HTTPUnit. Proc - Agil Confernce 2005 2005:29–38. doi: 10.1109/ADC.2005.11
10. Peeters J (2005) Agile security requirements engineering. … Requir. Eng. Inf. Secur.
11. Kitchenham B, Charters S (2007) Guidelines for performing Systematic Literature Reviews in Software Engineering. Engineering 2:1051. doi: 10.1145/1134285.1134500
12. Liu DLD, Wang QWQ, Xiao JXJ (2009) The role of software process simulation modeling in software risk management: A systematic review. 2009 3rd Int Symp Empir Softw Eng Meas 302–311. doi: 10.1109/ESEM.2009.5315982
13. Musa Shuaibu B, Md Norwawi N, Selamat MH, Al-Alwani A (2013) Systematic review of web application security development model. Artif Intell Rev 43:259–276. doi: 10.1007/s10462-012-9375-6
14. Usman M, Mendes E, Weidt F, Britto R (2014) Effort estimation in agile software development. Proc 10th Int Conf Predict Model Softw Eng - PROMISE '14 82–91. doi: 10.1145/2639490.2639503
15. Mozucha J, Bruno R (2008) Product-Focused Software Process Improvement. 5089:217–232. doi: 10.1007/978-3-540-69566-0
16. Cohn-Muroy D, Pow-Sang JA (2016) Trends and Applications in Software Engineering. Adv Intell Syst Comput 405:15–24. doi: 10.1007/978-3-319-26285-7

17. Kongsli V (2006) Towards agile security in web applications. In: 21st ACM SIGPLAN Symp. Object-Oriented Program. Syst. Lang. Appl. ACM Press, New York, New York, USA, pp 805–808
18. Wiesauer A, Sametinger J (2007) A SECURITY DESIGN PATTERN TAXONOMY Findings of a Systematic Literature Review. Proc SECRYPT 2009 Int Conf Secur Cryptogr 387–394.
19. Tomanek M, Klima T (2015) Penetration Testing in Agile Software Development Projects. Int J Cryptogr Inf Secur 5:01–07. doi: 10.5121/ijcis.2015.5101
20. Choliz J, Vilas J, Moreira J (2015) Independent Security Testing on Agile Software Development: A Case Study in a Software Company. 2015 10th Int Conf Availability, Reliab Secur 522–531. doi: 10.1109/ARES.2015.79
21. Felderer M, Fourneret E (2015) A systematic classification of security regression testing approaches. Int J Softw Tools Technol Transf 17:305–319. doi: 10.1007/s10009-015-0365-2
22. Tommy R, Mhaisekar M, Kallepally S, et al (2016) Dynamic quality control in agile methodology for improving the quality. 2015 IEEE Int Conf Comput Graph Vis Inf Secur CGVIS 2015 233–236. doi: 10.1109/CGVIS.2015.7449927
23. Azham Z, Ghani I, Ithnin N (2011) Security backlog in scrum security practices. 2011 5th Malaysian Conf Softw Eng MySEC 2011 414–417. doi: 10.1109/MySEC.2011.6140708
24. Nair S, De La Vara JL, Sabetzadeh M, Briand L (2013) Classification, structuring, and assessment of evidence for safety - A Systematic Literature Review. Proc - IEEE 6th Int Conf Softw Testing, Verif Validation, ICST 2013 94–103. doi: 10.1109/ICST.2013.30
25. Ghani I, Yasin I (2013) Software Security Engineering In Extreme Programming Methodology: A Systematic Literature. Sci Int 25:215–221.
26. Bozic J, Wotawa F (2014) Security testing based on attack patterns. Proc - IEEE 7th Int Conf Softw Testing, Verif Valid Work ICSTW 2014 4–11. doi: 10.1109/ICSTW.2014.58
27. Häser F, Felderer M, Breu R (2014) Software paradigms, assessment types and non-functional requirements in model-based integration testing. Proc 18th Int Conf Eval Assess Softw Eng - EASE '14 1–10. doi: 10.1145/2601248.2601257
28. Hummel M (2014) State-of-the-Art: A Systematic Literature Review on Agile Information Systems Development. 2014 47th Hawaii Int Conf Syst Sci 4712–4721. doi: 10.1109/HICSS.2014.579
29. Heck P, Zaidman A (2016) A systematic literature review on quality criteria for agile requirements specifications. Softw Qual J. doi: 10.1007/s11219-016-9336-4
30. Torrecilla-Salinas CJ, Sedeño J, Escalona MJ, Mejías M (2016) Agile, Web Engineering and Capability Maturity Model Integration: A systematic literature review. Inf Softw Technol 71:92–107. doi: 10.1016/j.infsof.2015.11.002

# SESSION 2

## SIMULATION AND OPTIMIZATION

# Creating Artificial Societies for Traffic Simulation: a tool using SUMO and JADE

Guilherme Soares, Zafeiris Kokkinogenis, and Rosaldo J. F. Rossetti

Artificial Intelligence and Computer Science Laboratory (LIACC)
Faculty of Engineering, University of Porto, Portugal (FEUP)
{guilherme.soares,pro08017,rossetti}@fe.up.pt

**Abstract.** The rapid ever-increasing population and urban activities have imposed a massive demand on Urban Transportation Systems (UTS). These systems were not prepared for such events, so traffic congestion and defective metropolitan systems were a direct consequence of this deficiency. The explosion of computing technology brought together expertise from different scientific and technical disciplines giving birth to new computing and communication paradigms. Taking advantage of the modelling and simulation technologies we have devised a framework that combines the characteristics of Multi-Agent System Development Framework, JADE, and the microscopic traffic simulator, SUMO, for the development and appraisal of multi-agent traffic solutions in contemporary traffic and transportation systems. Therefore we present to the community of researchers and practitioners a tool that can instantiate a heterogeneous Artificial Society (AS) of drivers immersed into a realistic traffic environment.

**Keywords:** Artificial Society, Intelligent Transportation Systems, Multi-Agent System, Microscopic Traffic Simulator, JADE, SUMO

## 1 Introduction

The rapid ever-increasing population and urban activities has imposed a massive demand on urban transportation systems. The main problem is that most of the urban areas were not prepared for such hasty development which led to weak and defective metropolitan transportation systems [7].

Efficient transportation systems are crucial to an industrialized society, therefore rapid and effective solutions for traffic congestion are needed to prevent its negative impact in the city's social and economic welfare. A way to address this issue is resorting to the use of modelling & simulation.

Transportation domain presents an inherent complexity. It involves diverse heterogeneous entities either in infrastructure or in behaviour, (e.g. vehicles, pedestrians), that can interact, reflecting social behaviours that go from coordination to competition. Moreover, the transportation domain presents a high degree of stochasticity and dynamicity especially when considered in an urban context. Thereby, using simulation and taking advantage of its characteristics we can test with several management solutions or even changes in the network more cheaply and faster. Such approach can provide us with the possibility of comparing studies between new infrastructures' designs or control algorithms without having to interfere in the real world.

Taking advantage of this simulation technologies a new generation of mobility systems, Intelligent Transportation Systems (ITS), arose and could be implemented or polished before being applied [12]. The ITS arise as the synergy between Information and Communication Technologies and Urban Transportation Systems, which include vehicles and networks that move people and goods. Traditionally, mathematical equations describe the drivers and pedestrians movements taking into account several flow restraints, they are used to tackle traffic related issues and to model them. According to this approach, the traffic problem was handled as a whole, and the solution was a product of the fulfilment of all trips.

The formalization of ITS concept is to be considered a great achievement by the transportation engineering community of practitioners and researchers. However, in the last few years the traffic and transportation domain has made a breakthrough in the way it is conceived. The explosion of computing technologies in the last decades brought together expertise from different scientific and technical disciplines giving birth to new computing and communication paradigms. A new type of systems called socio-technical arose from such mutual conjunctions where people and technology live in symbiosis. The transportation and, generally speaking urban domain, could not be impermeable to such revolution. Indeed, it proves to be a valid test-bed where such new social and technological paradigms can be applied.

A new concept has been coined to deal with this revolution, Future Urban Transport (FUT) Systems. Within FUT the notion of mobility system overcomes the ITS limitations, instead of focusing only on the simple processes of transportation of good and persons it becomes more conscious in terms of environment, accessibility, equality, security, and sustainability of resources [15]. People are placed as a central aspect, as well their preferences, of the urban systems, forcing architectures to become more adaptable and accessible to their needs. Therefore, new technologies and methodologies are necessary to track these new models, which leads us to our motivation.

### Motivation

Normally, in traffic solutions development, the use of a simulator is very straightforward related to traffic flow and junction management. Despite the many attempts and published papers, the solutions presented do not make full use of the concept of intelligent agents.

Additionally, the multi-agent systems paradigm has become recognized as a useful approach for modelling and simulating complex systems [13].

Keeping the above mentioned revolution in urban transportation in mind and guided by the need to design more human-centric economic and environmental solutions, a framework that generates an urban context, meaning a traffic network and its inherent heterogeneous population, is necessary so that analysts and designer can study, develop and evaluate their policies and strategies.

The aim of this paper is to present community of researchers and practitioners with a tool that can instantiate a heterogeneous Artificial Society (AS) of drivers immersed in a realistic traffic environment. Experimentations with such AS can help us study emergent mobility patterns and how information or knowledge can affect the drivers' decision-making process. The concept of AS can be used by traffic managers or government institutions as a test-bed for strategies or policies analysis towards a social-aware use of resources, or other optimization approaches.

In order to achieve the coupling of AS with traffic urban environment we will devise an integration of a microscopic traffic simulator for large-scale networks with an agent development platform framework, in order to support all the necessary elements of the Artificial Society of Drivers.

The remainder of this paper consists of the following. Section 2 briefly describes Artificial Societies in the traffic domain. Section 3 overviews the several components, utilities and tools of our framework. Section 4 provides insight into our implementation and agent architecture. Section 5 illustrates a usage example of our framework in a real-world scenario. Finally, Section 6 offers discussion and concluding remarks.

## 2 Artificial Societies in Traffic Domain

Agent-based computing is one of the powerful technologies for the development of distributed complex systems [24]. The growing interest in agent technology results from its inherent ability to decompose a system into multiple agents to achieve a global objective.

One great potential of Agent-based Modelling and Simulation (ABMS), is its application to simulate human societies. Traditional social simulation models were quite naive; they only supported homogeneous populations and sub-populations and thus, were unable to couple different social science disciplines.

First, Thomas Schelling [20] with the segregation model and then Epstein and Axtell with "Growing Artificial Societies" [8] made a breakthrough on how we can use agents for theory exploration or generation.

Epstein and Axtell [8] said: *"Fundamental social structures and group behaviours emerge from the interaction of individual agents operating on artificial environments under rules that place only bounded demands on each agent's information and computational capacity"*.

Not surprisingly, traffic systems are inherently linked to an urban context, which can be represented as a society, allowing agent technologies and Multi-agent simulation to cope with this field.

This artificial society is composed by an environment where a set of heterogeneous agents operate, obeying to defined behavioural rules (social rules) manifesting global dynamics.

Social systems are among the most complex systems in our world. Coleman [6] introduced the "Coleman boat", see Fig. 1, a representation of the process of human actions and social changes, meaning that macroscopic actions do not just cause change of the macro state of a society, but at the same time causes a change in the micro state of most or all of the individual beings.

Urban tissue can be represented as a society that presents organizational features. Thus, the concept of Agents and Agent-based Social Simulation can cope with transportation domain approaches to optimize and design systems.

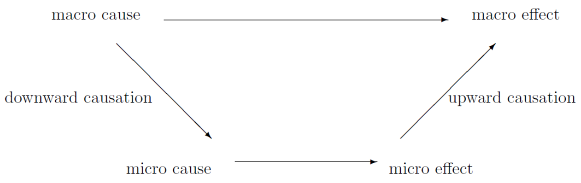


**Fig. 1.** The Coleman boat describing downward and upward causation [6]

As it is known, Metropolitan Systems are spatially spread and are composed by a great number of interacting entities in its domain, each of which with its individual behaviour. Herewith, the combination of all these specific actions have great influence on the overall system's condition. So, a good abstraction for these environments is agent-based traffic simulations, where each entity of the domain is modelled by an agent, e.g. vehicles, infrastructures, as well its behaviours.

We will focus our research in traffic simulation, specifically the Driver-Vehicle Agents. This term refers to the idea that a single entity, i.e. the agent, incorporates the complex human driving behaviour as well as the properties of a vehicle. The behavioural part of this entity reasoning is based on the decision-making process on a short or long term basis taking into account the characteristics of human behaviour in a given situation e.g. the quick overtaking decision or the re-routing in a network.

## 3 SUMO and JADE Overview

Having to deal with atomic entities in the transportation domain, i.e. vehicles or drivers, an Agent-Based Modelling approach seems to be the appropriate way to represent the road traffic environment and the driver entities that live and interact in it. A traffic simulator that can implement the agent paradigm must provide a good Application Program Interface (API) for accessing the vehicles' and other traffic variables, so that we can create, control and monitor its states from an external application.

We will also need an intuitive MAS development framework, which will be used to implement the drivers' artificial society. Thus it should be sufficiently portable, to match the simulator's execution environment and also be flexible enough to allow communication with the simulator's API.

The microscopic traffic simulator chosen to provide the environment and traffic entities simulation was SUMO, which is an open source, highly portable, microscopic traffic simulator. It provides real-time interoperability with other applications through the TraCIs API. This microscopic traffic simulator will provide us with the individual vehicular entities and the traffic environment.

We want to use a standardized and mature MAS development tool, so we chose JADE (Java Agent DEvelopment Framework). This free framework is highly portable since it has interoperating versions for every profile of Java Machines. JADE aims to simplify the development of MAS while guaranteeing standard compliance.

Our goal is not just to produce a framework for a single microscopic traffic simulator, but to any that fulfils the aforementioned requirements. So, instead of linking the MAS framework directly to the simulator's API, we will use TraSMAPI (Traffic Simulation Manager Application Programming Interface) as a generic microscopic traffic simulator API. This tool will allow a higher level of abstraction and transparency, which enables us to use any microscopic traffic simulator, allowing result comparison among other benefits.

### 3.1 SUMO

Simulation of Urban Mobility (SUMO) [2] is an open source, highly portable, microscopic and multi-model traffic simulation package designed to handle large road networks and to establish a common test-bed for algorithms and models from traffic research. The simulator was developed in the Institute of Transportation Systems at the German Aerospace Centre with collaboration of the Centre for Applied Informatics in Cologne and it is licensed under GPL.

SUMO is possibly the most studied open-source microscopic traffic simulator in the research community, with a high number of scientific papers referring to it. It is a complex project with several contributors [9] [14], and consists of hundreds lines of code, still growing. Its real-time interoperability is granted by TraCI's interface, which allows us to request data from an on-running traffic simulation.

SUMO is used in different projects in literature. It helps to investigate several research topics like simulating realistic Vehicular Ad-hoc NETworks (VANETs) [18], [17], to simulate automatic driving [16] or traffic management strategies, e.g. intelligent traffic lights [10], [11], route choice among others.

### 3.2 JADE

JADE (Java Agent DEvelopment Framework) is a free software framework to develop agent-based applications. Its goal is to simplify the development while ensuring standard compliance through a comprehensive set of system services and agents. JADE is fully implemented in Java language and is compliant with the Foundation for Intelligent Physical Agents (FIPA) specifications for interoperable multi-agent systems. Besides, this agent platform can be distributed across several machines, which do not even need to share the same Operative System (OS) [3].

The agent platform can be dispersed on several computers, where each of which runs a single Java Virtual Machine (JVM). Each JVM is a container of agents that provides a complete run time environment for agent execution and allows several agents to concurrently execute on the same host

Numerous R&D projects, where an interaction between several elements is required, and in which an autonomous and dynamic adaptation to complex relations is needed, have used JADE as a developing tool. In traffic domain, there are several works that profit from JADE platform for developing Multi-Agent System traffic management solutions [19], [22].

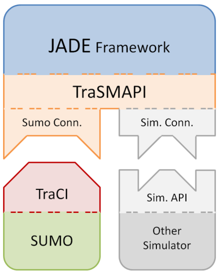


**Fig. 2.** Framework Architecture

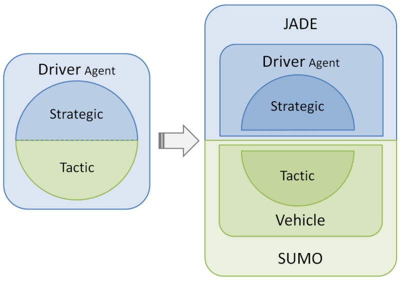


**Fig. 3.** Driver's Layers: tactic-reactive in SUMO; strategic-cognitive layer in JADE

### 3.3 TraSMAPI

TraSMAPI (Traffic Simulation Manager Application Programming Interface) is a synergy between two main components: an Application Programming Interface (API) and a Multi-Agent System framework [21].

The API was built in an abstraction level higher than most common Microscopic Traffic Simulators so that, ideally, the solution should be independent of the microscopic simulator choice. This is guaranteed as far as the chosen simulators allow it, and provided that their communication interface differs and they do not implement the same set of features.

This feature allows the comparison of results from different simulators using exactly the same traffic management solution.

The Multi-Agent System framework is a module that is meant to serve as a starting point for the creation of Multi-Agent Systems. It allows the creation of new agents by following a common interface. The agents themselves are created with a reference to one or multiple objects in the simulation gaining direct access to its artefacts or entities.

Concerning our work, we aim to substitute this MAS Framework Module for a more widely distributed MAS frameworks - JADE. With this, we orient TraSMAPI also to real-world solutions implementations, since it will have a more mature, generic and FIPA compliant MAS development framework.

## 4 Proposed Implementation

Our goal is to provide community with a tool that can instantiate a heterogeneous Artificial Society (AS) of drivers immersed into a realistic traffic environment. Accordingly, we have an artificial society of drivers in JADE platform, each of which responsible for one vehicle in the SUMO's traffic environment.

Taking into consideration all the general requirements and goals we have devised the following architecture, depicted in Fig. 2

In Fig. 2 we can observe the main contribution of TraSMAPI in our framework. TraSMAPI provides an abstraction over different possible microscopic simulators, which completely makes our platform independent from a microscopic traffic simulator. Besides, it makes possible further studies on comparing simulation results. Since this architecture enables us to test the same solution, i.e. source code, in various microscopic simulators, one can analyse the possibility of different outcomes.

The microscopic traffic simulator offers an API for access to its simulation state - TraCI. For an external application to communicate with this software it must obey the TraCI communication protocol and messages types. The Sumo Communication Module attached to TraSMAPI, converts this low-level simulator's API to a higher-level one, which will be then used by our artificial society of drivers implemented in JADE's MAS development framework coupled to TraSMAPI.

### Agent Architecture

To build and associate each driver agent with a simulated vehicle and endorse him with all the driving decisions, skills and cognitive characteristics would be computationally very expensive. To simulate hundreds or thousands of vehicles and driver's decision-making in JADE we have adopted the delegated-agent concept, which has been used in [23], to separate the tactical from the strategic layer of the agent, and execute them in parallel, thus improving performance, Fig. 3.

The tactic-reactive layer was entrusted to the microscopic traffic simulator, taking care of reactive tasks related to driving itself e.g. break, change lane, slow down, overtake. The strategic layer was kept in JADE's framework, allowing researchers to implement complex agent architecture, from pure reactive to pure cognitive. This layer, is intended for more demanding and elaborate reasoning like re-routing, collect or spread traffic information among other entities.

## 5 Experimental Set Up and Results Discussion

To illustrate the capabilities of the platform, our experimentation scenario is related to the Braess Paradox in a route choice scenario, where the Artificial Society's capability of learning in adverse and dynamic situations emerges.

There are generally two types of travel behaviour: user-optimizing behaviour, in which travellers are generally characterized as selfish and therefore select their optimal route; and system-optimizing behaviour, in which a central controller directs traffic. Our work focuses on the first one, the driver behaviour, and the Braess paradox occurs only for user-optimizing behaviours.

### Braess Paradox

In an urban area with a lot of traffic, adding a new road to distribute and facilitate traffic may seem an intuitive idea. However, according to the Braess paradox, just the opposite occurs: a new route added in a transportation network actually increases the travel time of all individual travellers[4, 5].

The Braess Paradox is a good illustration of how easily our intuitions about collective interaction can be fooled.

Car drivers seek to minimise the time to get from O (Origin) to D (Destination), however, car drivers may not be able to act independently of each other: collective interactions may influence individual behaviour.

We have made this experimentation scenario as a proof of concept of one of the numerous uses that this platform provides to the community of researchers and practitioners. In this case we tried to replicate the Braess paradox by setting up an artificial society of "selfish" learning drivers, in a well defined scenario. Their goal is to get from point O to point D the fastest way possible.

The network, sketched in Fig. 4, starts by being composed by two symmetrical routes, each of which consists of a fast section and a slow one. Then, a new fastest road is added, Fig. 5, providing drivers more and better road resources. This network change could be considered as a new bridge construction possibility in the real-world scenario, allowing a fastest way into the cities.

We have built an artificial society of Q-learning drivers, which will "live" for 500 days and perform, each day, a trip from point O to D. When arrived, each driver registers his Travel Time (TT):

$$TT = arrivalTime - departureTime \tag{1}$$

Taking the environment into account we have modelled it in a finite-state automaton, with 3 edges from node 'O' to node 'D', and we have built the correspondent Q-table to each of the driver agents, where each route choice in state *s* generates an *utility*.

Since our problem is scalar, depending only on the route choice and not from the current state, we can simplify it to *Q(r)*, being *r* the route chosen. Hence our utility-function is:

$$Q(r) = (1-\eta)\,Q(r) + \eta.R \tag{2}$$

being $\eta$ the learning rate and $R$ the Reward function:

$$R = \frac{aTT}{TT} - 1 \tag{3}$$

while *aTT* is the average Travel Time of all trips:

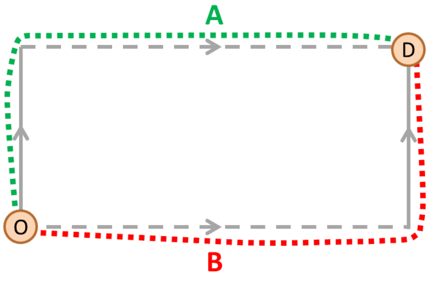


**Fig. 4.** Two route network

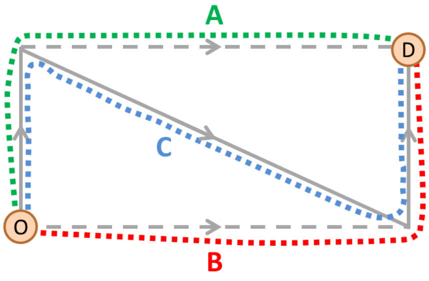


**Fig. 5.** Three route network

$$aTT = \frac{\sum TT}{trips} \tag{4}$$

For our test-bed we have defined an exploring and an exploiting time in each network configuration for all 500 days.

Each network configuration, meaning different route arrangement, is **explored** by the driver agents during 50 days, in which the drivers are randomly assigned to a route so to retrieve knowledge from its journey time, thus updating his Q-table. The remaining days are **exploited** by the driver according to its utility values. The drivers' departure time is equally distributed along the first hour of the day. So in the two-route scenario (Fig. 4) the drivers will perform 50 day of exploration and 150 days of exploitation. Afterwards, they will have another 50 days of exploration in the three-route scenario (Fig. 5) and 250 days for exploiting their best options.

We have performed several tests with various number of drivers to observe their learning process in a route-choice environment. We have managed to extract two different patterns: On the one hand we have made experimentations with a very low vehicle density, approximately 1 vehicle each five seconds, which did not evidence the paradox scenario. Since the departure times were very temporally spaced, the new route was not jammed and, therefore chosen by the majority of drivers.

On the other hand we have explored a test-bed with a high density of vehicles, approximately 1 vehicle each two seconds. In this case, we have noticed the increase of the travel time and the **underutilization** of the newly added route, despite being the fastest one. This experiment is plotted if Fig. 6

**Results Discussion**

In the first **exploitation** phase [50,200], the number of vehicles that chose route A or route B is nearly the same, without fluctuations, which establishes a constant average travel time (observed in the bottom graph).

During the second **exploration** phase [200,250], we verify that the average travel time in the new route C is a bit smaller. Hence, in the beginning of the second **exploitation** period [250,500] the drivers should have a great utility in the choice of C. In fact, we can observe that almost every 1900 vehicles chose to travel through it i.e. route C, **over exceeding the initial average travel time**, recorded when there were only 2 routes available. With this insertion, the average travel time increased from approximately 1000 seconds to a staggering 3000 seconds.

The learning drivers, encountering such a scenario, quickly change their opinion on the utility of route C. They return to their previous choice avoiding the overpopulated route and improving their travel time. We can observe this event in the quick variation in the upper graph's peaks in just approximately 20 days.

With this learning process the overall travel time diminishes as well as the **underutilization** of route C, which becomes the less used route, despite being the fastest one.

Our main purpose is not to discuss the Braess Paradox, as it has already been done by Bazzan & Klügl in [1]. However, with this tool, we could perceive the emergent social behaviour which is the Braess Paradox. Hence, we could illustrate how the tool

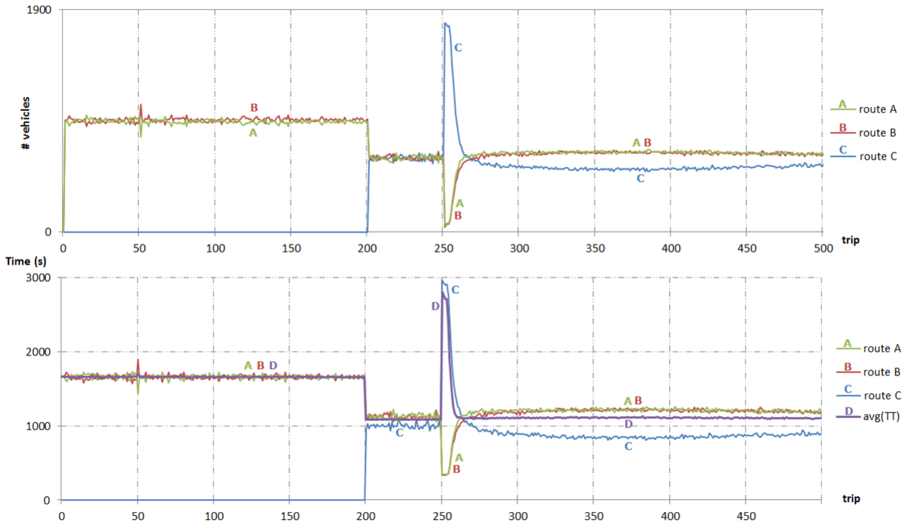


**Fig. 6.** Occupation and Travel Time in the 1900 Vehicle Test.
Top: Occupation(trip); Bottom: Travel-Time(trip)

can be used in a realistic case study, using a microscopic traffic simulator, having an AS of learning drivers implemented to represent knowledge acquisition and exploitation of a network.

## 6 Conclusions

Traffic systems have been subject to a lot of improvement last decades and travellers have, in general, witnessed a revolution in the way a trip is planned in urban networks. Hence, facing the current traffic situation in most developed countries it is now imperative to foster new transportation methods using state-of-the-art technologies towards Future Urban Transport (FUT).

Simulation proved to be an effective approach to analyse and design novel traffic solutions in socio-technical aspect systems. We have devised a conceptual architecture and we built a prototype of Artificial Societies in traffic scenario, where we experimented the knowledge representation of the network using Reinforcement Learning techniques, using the concept of delegate agent, where the agent uses tactic and strategic layers for reasoning.

We proposed a framework where MAS of different nature can be instantiated over the traffic domain, meaning socio-technical systems, embedded intelligent artefacts, aiming to design more human-centric, economic and environmental solutions. This tool also reveals great flexibility for multi-agent systems design and development in traffic domains, since the developer can easily model and test his own synthetic population by implementing their artificial society of drivers architectures, where each agents is presented with its own preferences and beliefs. Such AS can thus be used to design solutions based in individual or collective intelligence and participation (social-awareness) or as a test-bed for policy evaluation by governmental institutions.

As future developments, not only, inter-vehicular or vehicle-to-infrastructure communications, but also, new policies development and incentive mechanisms study, could be built and appraised.

## References

1. Bazzan, A.L.C., Klügl, F.: Case studies on the Braess Paradox: Simulating route recommendation and learning in abstract and microscopic models. Transportation Research Part C: Emerging Technologies 13(4), 299–319 (Aug 2005), http://www.sciencedirect.com/science/article/pii/S0968090X05000331
2. Behrisch, M., Bieker, L., Erdmann, J., Krajzewicz, D.: SUMO - Simulation of Urban MObility: An Overview. In: SIMUL 2011, The Third International Conference on Advances in System Simulation. pp. 63–68. Barcelona, Spain (2011)
3. Bellifemine, F., Poggi, A., Rimassa, G.: JADEA FIPA-compliant agent framework. In: Proceedings of PAAM. p. 33. London (1999)
4. Braess, D.: Über ein Paradoxon aus der Verkehrsplanung. Mathematical Methods of Operations Research 12(1), 258–268 (1968)
5. Braess, D., Nagurney, A., Wakolbinger, T.: On a paradox of traffic planning. Transportation science 39(4), 446–450 (2005)
6. Coleman, J.S.: Foundations of social theory. Belknap Press (1994)
7. Doytsher, Y., Kelly, P., Khouri, R., McLAREN, R., Potsiou, C.: Rapid urbanization and mega cities: The need for spatial information management. Research study by FIG Commission 3. FIG Publication No 48 (2010)
8. Epstein, J.M., Axtell, R., Project, .: Growing Artificial Societies: Social Science from the Bottom Up. Complex Adaptive Systems, Brookings Institution Press (1996), http://books.google.pt/books?id=8sXENe8QrmYC
9. Flotterod, G.: Cadyts–a free calibration tool for dynamic traffic simulations. In: Proceedings of the 9th Swiss Transport Research Conference (2009)
10. Krajzewicz, D., Brockfeld, E., Mikat, J., Ringel, J., Rössel, C., Tuchscheerer, W., Wagner, P., Wösler, R.: Simulation of modern Traffic Lights Control Systems using the open source Traffic Simulation SUMO
11. Macedo, J., Soares, M., Timoteo, I., Rossetti, R.J.F.: An approach to advisory-based traffic control. Information Systems and Technologies (CISTI), 2012 7th Iberian Conference on pp. 1–6 (2012)
12. Miles, J.C., Walker, A.J.: The potential application of artificial intelligence in transport. Intelligent Transport Systems, IEE Proceedings 153(3), 183–198 (2006)
13. Moya, L.J., Tolk, A.: Towards a taxonomy of agents and multi-agent systems. In: Proceedings of the 2007 spring simulation multiconference - Volume 2. pp. 11–18. SpringSim '07, Society for Computer Simulation International, San Diego, CA, USA (2007), http://portal.acm.org/citation.cfm?id=1404682http://dl.acm.org/citation.cfm?id=1404680.1404682
14. Papaleondiou, L.G., Dikaiakos, M.D.: Trafficmodeler: A graphical tool for programming microscopic traffic simulators through high-level abstractions. In: Vehicular Technology Conference, 2009. VTC Spring 2009. IEEE 69th. pp. 1–5. IEEE (2009)
15. Passos, L.S., Rossetti, R.J.F., Kokkinogenis, Z.: Towards the next-generation traffic simulation tools: a first appraisal. Information Systems and Technologies (CISTI), 2011 6th Iberian Conference on pp. 1–6 (2011)

16. Pereira, J.L.F., Rossetti, R.J.F.: An integrated architecture for autonomous vehicles simulation. In: Proceedings of the 27th Annual ACM Symposium on Applied Computing. pp. 286–292. SAC '12, ACM, New York, NY, USA (2012), `http://doi.acm.org/10.1145/2245276.2245333`
17. Piórkowski, M., Raya, M., Lugo, A.L., Papadimitratos, P., Grossglauser, M., Hubaux, J.P.: TraNS: realistic joint traffic and network simulator for VANETs. SIGMOBILE Mob. Comput. Commun. Rev. 12(1), 31–33 (2008), `http://doi.acm.org/10.1145/1374512.1374522`
18. Rieck, D., Schünemann, B., Radusch, I., Meinel, C.: Efficient traffic simulator coupling in a distributed V2X simulation environment. In: Proceedings of the 3rd International ICST Conference on Simulation Tools and Techniques. pp. 72:1—-72:9. SIMUTools '10, ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering), ICST, Brussels, Belgium, Belgium (2010), `http://dx.doi.org/10.4108/ICST.SIMUTOOLS2010.8640`
19. Sanchez Passos, L., Rossetti, R.: Traffic light control using reactive agents. In: Information Systems and Technologies (CISTI), 2010 5th Iberian Conference on. pp. 1–6. IEEE (2010)
20. Schelling, T.C.: Dynamic models of segregation. Journal of mathematical sociology 1(2), 143–186 (1971)
21. Timoteo, I.J., Araujo, M.R., Rossetti, R.J., Oliveira, E.C.: TraSMAPI: An API oriented towards Multi-Agent Systems real-time interaction with multiple Traffic Simulators. 13th International IEEE Conference on Intelligent Transportation Systems pp. 1183–1188 (Sep 2010), `http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=5625238`
22. Van den Bosch, A.T., Menken, M.R., van Breukelen, M., van Katwijk, R.T.: A test bed for multi-agent systems and road traffic management. In: Proceedings of the 15th Belgian-Netherlands Conference on Artificial Intelligence (BNAIC'03). pp. 43–50 (2003)
23. Wahle, J., Bazzan, A.L.C., Klügl, F., Schreckenberg, M.: The impact of real-time information in a two-route scenario using agent-based simulation. Transportation Research Part C: Emerging Technologies 10(56), 399–417 (Oct 2002), `http://www.sciencedirect.com/science/article/pii/S0968090X02000311`
24. Zambonelli, F., Van Dyke Parunak, H.: Signs of a revolution in computer science and software engineering. Engineering Societies in the Agents World III pp. 120–125 (2003)

# Rush Hour Traffic Conditions Impact in Electric Bus Performance: A Case Study in Porto

José Macedo and Rosaldo J. F. Rossetti

Artificial Intelligence and Computer Science Laboratory (LIACC), Department of Informatics Engineering (DEI), Faculty of Engineering, University of Porto, Portugal
{jose.macedo,rosaldo.rossetti}@fe.up.pt
http://www.fe.up.pt

**Abstract.** The increasing traffic volumes inevitably leads to traffic congestion, affecting the economic activity of cities and resulting in heavy environmental condition. One of the solutions under analysis is the development of electric vehicles for personal and public transportation. However, there still remains some concerns related to the consumption of energy for considering this implementation as a cost-effective solution. For instance, the fact these vehicles are immersed in an urban context is an important aspect in evaluating its performance and effectiveness. This research follows an High-Level Architecture (HLA) based distributed architecture for electric bus powertrain simulation within a realistic urban mobility environment, to analyze the performance of an electric bus for two different routes in the city of Oporto under a rush hour traffic flow condition and a free-flow basis. The intent is to study the influence of rush hours traffic dynamics in the amount of energy spent on the routes courses. To this effect, a simulation platform was built coupling the SUMO microscopic traffic simulator with a model of an electric bus powertrain subsystem designed in the MatLab/Simulink environment. Simulation results are compared and discussed, with one of the key conclusions being that the performance of the electric bus is negatively affected by the rush hour traffic conditions, and that it is more evident in a more linear route topology.

**Keywords:** Simulation Platform, Electric Vehicle Performance, Public Transport, Route Planning, Traffic Flow Dynamics

## 1 Introduction

Over the last decades we have seen an increase in traffic and transport demand that has intensified capacity problems in the infrastructure causing traffic congestions and delays. Even with the constant expansion of the street and road system in large cities, traffic congestion continues to be one of the major transport issues in practically every urban area. According to the 2014 Urban Mobility Report [14] the congestion led urban Americans to travel 6.8 billion hours more in 2013, which resulted in 3.1 billion gallons of extra fuel, and delay will grow to 8.2 billion hours and wasted fuel will increase to 3.7 billion gallons in 2020.

Traffic congestion can have several effects on the economy, the environment and people. Problems in the traffic system have a huge impact on virtually all areas of economic activities, considering the flow of people and goods between cities is directly related to the road network [1]. Additionally, the time and fuel wasted while citizens are stuck in traffic not only affect their financial lives but can also lead to serious health issues, either due to stress, or to high air pollution levels. Besides, personal vehicles are a major cause of global warming.

There is, of course, a diversity of different solutions trying to tackle congestion problems. As congestion begins to occur when the amount of traffic on a road network is approaching its maximum capacity, and the most obvious solution is to increase the network capacity [16]. This can be done in several ways, such as building new roads, extending the existing ones and adjusting the speed limit of roads to increase their capacity. However, creating new roads or adding additional capacity to the existing ones can be expensive, time consuming, can cause environmental and social impacts and is not guaranteed that it solves the problem, as demonstrated by the Braess' paradox [2]. Thus, the increase in traffic volumes combined with often short distances between intersections requires the adoption of a systems analysis approach to properly address traffic congestion. Often, traffic congestion is not the result of excessive traffic, but the result of overlapping bottleneck locations. The spillover effect of traffic congestion from one location to another makes conventional engineering methods inefficient [11]

Electric vehicles seem to offer a sustainable solution since they use electricity as fuel, producing less emissions than the conventional ones powered by fossil fuels, and when the electricity is produced from renewable sources, electric vehicles produce zero emissions [15]. On the other hand, public transportation is crucial to a citys transportation system and is essential to the quality of life of citizens. Combining these two important benefits, one can have the implementation of electric buses in metropolitan transportation as a sustainable solution.

Although buses offer a more flexible and satisfactory service to a larger number of citizens than other public transportations such as trolleys, subways or trains, planning routes for buses is a true complex matter. A public transportation system must provide a large number of individual trips. Besides, it is important to offer a network of routes, so that passengers can use a combination of routes to reach their destinations. Furthermore, the transportation systems should be parts of coordinated networks rather than routes planned in isolation, and must provide suitable links between all points where there is a demand.

But in the case of electric buses it is essential to take into account other important factors, such as the total energy of the buses batteries and the topology profile of each route and the urban environment context, which has a large influence on the bus performance and autonomy. In fact, there are still open issues related to the consumption of energy and other performance measures for considering the adoption of electric buses in urban scenarios as a cost-effective solution [9].

There is a general lack of evidence in the literature related to tools and methods that can accurately assess the performance of electric buses in urban scenarios. Thus, this paper will focus on the investigation of the amount of energy required to complete two standard bus driving cycle routes under rush hour traffic flow conditions, in order to evaluate the influence of traffic dynamics in the electric bus performance. It could be useful to understand the importance of urban context in the electric bus route assignment process.

This is accomplished by the use of an integrated simulation platform, following the High-Level Architecture (HLA) approach for distributed simulation. This platform accounts for the microscopic traffic simulator SUMO (Simulation of Urban MObility) which was coupled with a model of an electric bus powertrain (EBPS) designed in the Matlab/Simulink environment.

The remaining part of this paper is organized as follows: Section 2 approaches to the related work concerning the main subject of this paper; Section 3 describes all components of the integrated simulation framework; Section 4 defines the methodological approach; Section 5 describes the experimental setup made for the simulations and section 6 stresses some accomplished results and discussion; For last, section 7 concludes with final remarks and future work.

## 2 Related Work

Related work on traffic conditions impact in electric vehicles performance covers other works that propose an analogous approach to study this subject. The first issue to overcome is to find a way to simulate an electric vehicle performance on urban environment, considering the realistic representation of the route.

*Deborah Perrotta et al.* presented a previous approach to this problem in [12]. In this paper was studied the correlation between the type of route a bus travels and the amount of energy it spends to do it. In addition to energy, in this study was also discussed the potential of the recovering part of it through regenerative braking.

A suggestion for a simulation framework for electric vehicles in terms of energy consumption, offering a means to perform studies of electric mobility in urban areas, can be found at [10]. This tool extends the SUMO traffic simulator in a 3D simulation environment and uses a component that represents an electric vehicle coupled to the proposed car-follower model [10]. However, this platform is entirely built on top of the modules added to SUMO, unlike the proposal presented in this paper that integrates different simulators.

Although, [7] presents a platform that integrates Sumo and Simulink to simulate wireless communication between vehicles. This integration is accomplished by converting the model implemented in Simulink to C++, once SUMO is C++-based [7]. In this integrated platform, Sumo was responsible for a vehicle mobility behaviour whose traffic mobility-related parameters were given by its controller supplied by the Simulink model. *Chalé Góngora et al.* propose a system which generates a smart energy plan for the driver in [3]. They used the Generalized Multi-Commodity Network Flow (GM-CNF) system as a vehicle routing model that studies energy consumption and charging time to assure the use of an electric vehicle beyond its autonomy by choosing the best routes to arrive the place of destination with minimal time and cost.

The mentioned references propose integrated platforms to solve problems related to energy consumption and electric mobility in urban areas, but none of them considers the traffic conditions effect in the electric vehicle performance. Besides, the second reference does not integrate different simulators and the third one integrates two simulators in a different way. Therefore, the study herein presented has a vast potential, both with respect to the interoperability and the reuse of the software, as well as in the analysis of the influence of heavy traffic on the energy consumption of the vehicle.

## 3 Framework Architecture Concepts

### 3.1 HLA Concepts

Approaching the reuse and interoperation of the legacy model simulations, the High Level Architecture concept attend to a distributed simulation considering that no single simulation model can satisfy the requirements of all usages and users. Therefore, HLA provides easier interoperability and reusability discerning the simulation functionality provided by the members of the distributed simulation from a set of basic services for data exchange, communication and synchronization.

In HLA, every participating application is called a *federate*, and these entities can interact with each other within a *federation*. A federation can be seen as a set of federates acting together in a distributed simulation to achieve a certain objective. There are three main components that comprise HLA:

- Federate Interface Specification
- Framework and Rules
- Object Model Template Specification

The HLA Framework defines a set of rules that must be obeyed to ensure the proper interaction of federates within a federation. These rules must be unchanged across all the simulation units as they define the overall architecture. They also define the responsibilities of federates and federation. There are five rules for federates and other five for federations. The definition and description of each rule is available in [5].

The HLA Federate Interface Specification describes the services which federates have to use for communicating with others. This communication is always made through a middle-ware structure, known as Run-Time Infrastructure, which provides the essential building ground for the software developers. The interface specification describes which services a federate can use and which services it has to provide [4]. In order to establish the interaction between federates and the Run-Time Infrastructure (RTI), the concept of ambassador is used. Ambassadors are objects that have the methods needed by the participants for performing communication. So, federates communicate with the RTI using its ambassador as an interface.

The HLA Object Model Template Specification describes the format and syntax of the data transferred between federates. This data exchange is represented in the form of an object class and the two types of object exchange are Object Class and Interaction Class. The first one contains the shared information within federation that persists during the run time. The second one, contains the sent and received information between federates. This component defines the object template data that all simulation unit needs to use in order to exchange data with each other [6].

### 3.2 SUMO-SIMULINK: An HLA Integration

With the purpose of overcome the absence of appropriate tools to evaluate the performance of electric vehicles on urban areas, considering the detailed representation of the routes and urban traffic conditions, an integrated simulation platform was implemented as it has been presented in [8].

The platform accounts for the integration two different systems. The first one is concerned to the traffic system representation, which includes the road network (the physical infrastructure and the topology) and the traffic flow dynamics (the vehicle-entities and its interactions). The other, represents the electric bus system defined in terms of its powertrain subsystem such as the set of battery and traction motor, among others.

The systems integration is achieved by associating the electric bus powertrain subsystem to a vehicle entity (corresponding to a vehicle of class bus) of the microscopic traffic model.

With the intention of implement the physical road infrastructure and the traffic dynamics through vehicular movements in microscopic level resolution, the SUMO software suite has been considered. And for the simulation of electric bus operations and performance, a mathematical model of an electric bus implemented in Simulink has been considered [13]. Fig. 1 illustrates the main components of the integration architecture using the HLA concepts.

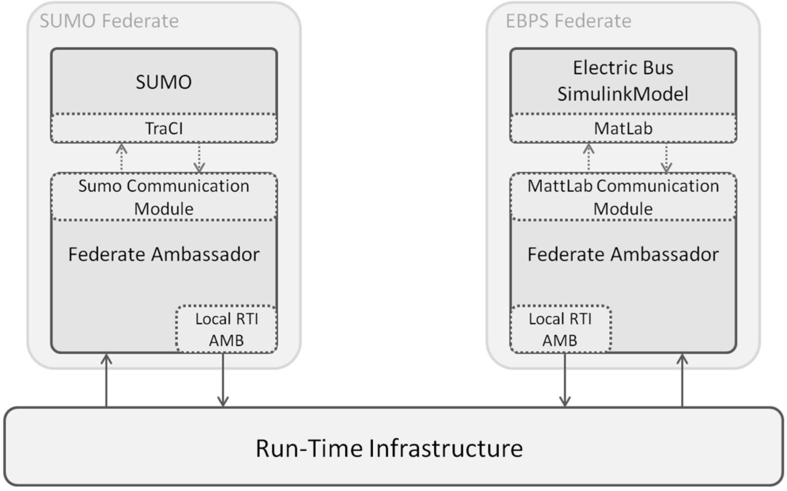


**Fig. 1.** HLA Implementation Architecture. [8]

**The MatLab Communication module** MatLab could be seen as an API for Simulink since Simulink models can be controlled by MatLab methods. Simulink models are standalone models and exist in mutual cooperation with the MatLab environment. The only way to access them externally is through the MatLab methods and calls. For this reason, was applied a control using the MatLab interface to Simulink. In this direction, this module is used by the federate ambassador to communicate with Simulink model through MatLab.

With the purpose of establish communication with MatLab, the matlabcontrol API was used. The matlabcontrol is a Java API to interact with MatLab letting its methods invoke behaviour of Java objects [14]. Using this, it was only necessary to create a proxy to work as an image of the MatLab application. After this procedure, all the calls to MatLab are performed through this proxy.

Implementing the communication with MatLab was the first stage in the development of this module. The second one was to control the Simulink model's step-by-step simulation through Matlab commands. For this purpose, the set_param() function was chosen. This function allows starting, pausing, stopping and restarting the simulation, as well as advancing one simulation step at a time.

**The Sumo Communication Module** Contrasting Matlab/Simulink, SUMO already comes with an API that offer a communication protocol. In fact, the TraCI interface provides a set of methods that permit an easy interaction with the simulators state variables.

This module is compounded of a set of functions, implemented in Java that include the indispensable commands that would be used by the SUMO federate ambassador to interact with SUMO. TraCI has a considerable number of methods, each one associated with an entity in the simulation. For the extent of this project, the only TraCIs methods that were used are relative to the vehicle entity and its speed.

**Specification of Federates** There are two different groups of methods that are related to the type of data exchange. The first group is directed to the interaction classes and the other, to the object classes. But before exchanging data between federates, the Federation Object Model needed to be specified and the communication module with the simulators needed to be created. See [8] for a detailed specification of Federation Object Model (FOM).

Related to the interaction classes, a set of methods were used in order to allow one of the federates to interact with the other. These methods acknowledge what kind of interaction the federates are able to publish or receive. For example, if a federate will be responsible to instruct the simulation to start, this federate could publish an interaction class named ”start” while the others subscribe it. In this case, the SUMO federate will be the responsible to initiate the simulation and to publish an interaction class named ”Start”. The EBPS federate will need to subscribe that interaction in order to receive it.

In a similar way each federate needs to call ”Publish” or ”Subscribe” for the *ObjectClass* and *ObjectClassAttributes* that they want to send or receive. A further description on federates object classes specification and the necessary methods for exchanging data during federation execution could be seen in [8].

## 4 Simulation Scenario

As previously mentioned, the integrated simulation platform that has been presented in [8] was used, in order to evaluate the performance of electric buses on urban environment, accounting for the detailed representation of the routes they would perform.

In this section, a scenario of two different routes are specified. Each one of the routes are performed considering two different traffic flow conditions: one on a free-flow basis and other during the rush hour. With this scenario, we intend to analyse the performance of the EBPS concerning the total energy required to complete each route in different day periods.

For this purpose, a road network has been created, a Origin/Destination matrix has been designed and the two bus routes and their bus stops were specified to set the test scenario in the simulator SUMO.

### 4.1 Network

The road network used in this test-bed was a model of the city of Porto. To create the network model, it was used the *OpenStreetMaps*[1] database to extract the preliminary information of the city's road network. The *OpenStreetMaps* is a collaborative project to create a free editable map of the world that provides useful data for a good starting point in the network modelling process. However, the extracted information was not completely correct and lacks on some important information for the simulation such as turnings possibilities at the intersections. Therefore, a meticulous adjustment of the number of lanes, maximum speeds and bus lanes for each road, as well as the specification of the turns allowed at each intersection was made using the capabilities of *PostGIS*[2] software over a geographic database and the *QGIS*[3] as viewing and editing tool.

### 4.2 Routes

Regarding the bus routes, the choice was made based on some specific characteristics such as route size, number of stops, distance between stops and the zones affected. Therefore, it was chosen the routes presented in Fig. 2, corresponding to the Lines 305 and 302 respectively.

Line 302 is a regular urban route that is comprised in the centre of city. This route is characterized for having some narrow streets and more irregular topology. Line 305 is more linear than line 302, with longer straight roads and it covers a larger area starting at the centre area of Porto down-town and ending at the outskirts of city. The routes have approximately the same size, about 8.1 km, and have almost the same number of bus stops, that is 23 and 26 respectively.

[1] http://www.openstreetmap.org/

[2] http://www.postgis.net/

[3] http://www.qgis.org/en/site/

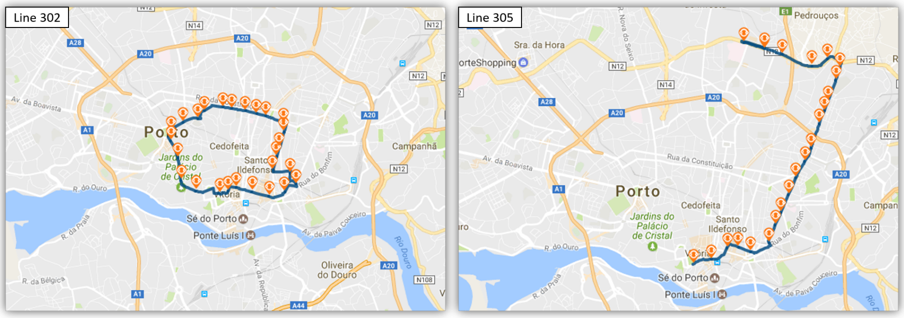


**Fig. 2.** Bus Line routes for test-bed experiments.

### 4.3 Traffic Flow Conditions

For the scenario where an ideal traffic flow was considered, it was not included any traffic demand in the simulation allowing the electric bus to move on free-flow roads. In other hand, to simulate the traffic dynamics during the rush hour a four-step model basis process was conducted to produce the necessary data.

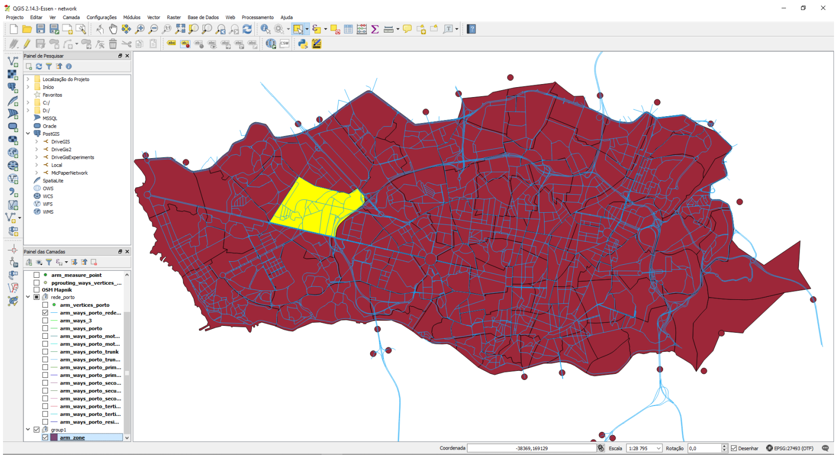
The Four-Step Models composed by the following steps:

- Trips Generation
- Trips Distribution
- Modal Split
- Route Assignment

The two first steps include the specification of zones, the number of trips each zone generates and how the trips are distributed to the other zones. In order to do that, the results from a study carried out by TIS (Transportes, Inovação e Sistemas) for the CMP (Câmara Municipal do Porto) were made available to the authors. This data contains the geometries of the zones and an Origin/Destination matrix of the number of trips from zone to zone during the morning rush hour. Importing the zones to the the geographic database and using the capabilities of *PostGIS*, it was possible to associate each road to a specific zone, Fig. 3 shows the *QGIS* visualization of the roads and zones.

Regarding the modal split step, there was no need to split the trips through different transport modal as the trips only accounts for private cars. Saying that, we consider that all trips are from private cars in the simulation. The last step was to assign the trips to routes. To do so, the *Duarouter* application embedded to SUMO package was used to automatic generate the routes for the vehicles to be used in the simulation. Fig. 4 shows the flow of operations to draw the necessary input files to perform the simulation.

As can be seen, the simulator SUMO needs two files as input for the simulation: one is the *Net* file which represents the road network infrastructure, such as the roads, its shapes, maximum speeds, number of lanes, among other information; the other is the *Routes* file that represents the traffic demand, it includes the route of each vehicle that will participate in the simulation and its entrance time.



Coordenada -38369,169129 Escala 1:28 795 Rotação 0,0 ☑ Desenhar EPSG:27493 (OTF)

**Fig. 3.** Zones and roads in QGIS graphical user interface

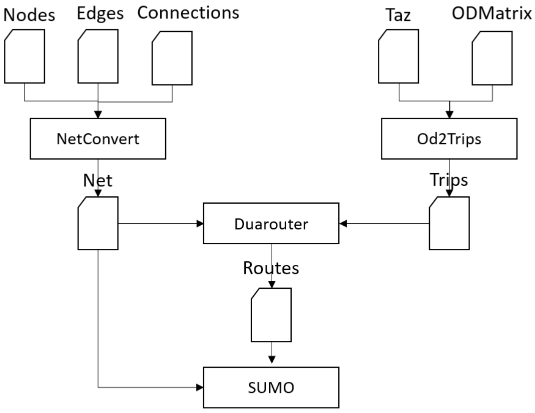


**Fig. 4.** Operational flow of creation of the inputs for the simulation.

The network input file was created using the *Netconvert* tool from SUMO package. This tool requires nodes description, edges specification and connections between the edges to generate the network file. Regarding the *Routes* file, it was obtained using the *Duarouter* application from SUMO package, as previously mentioned. This application needs the *Net* and *Trips* files as input. The Trips file contains all trips information specifying the depart time, depart edge and arrive edge. This file is generated using the *Od2Trips* tool, also included in SUMO package. This tool requires *Taz* file, which includes zones description and the specification of source and sink edges for each one, and the *ODMatrix* file that contains the trips distribution represented by an origin/destination matrix.

## 5 Experimental Results and Discussion

Four simulations were performed for the analysis of the electric vehicle performance in terms of energy consumption: one per each route with traffic demand and other per each route with no traffic at all. Some data possible to collect included speed, power, energy and braking resistance energy. To this analysis the most important parameter to take into account is the amount of consumed energy.

After running the simulation scenarios it was possible to calculate the amount of energy that each driving cycle requires for being completed. Fig. 5 shows the total energy that the electric bus spent to complete each route under the intense traffic flow of the rush hour and on a free-flow base.

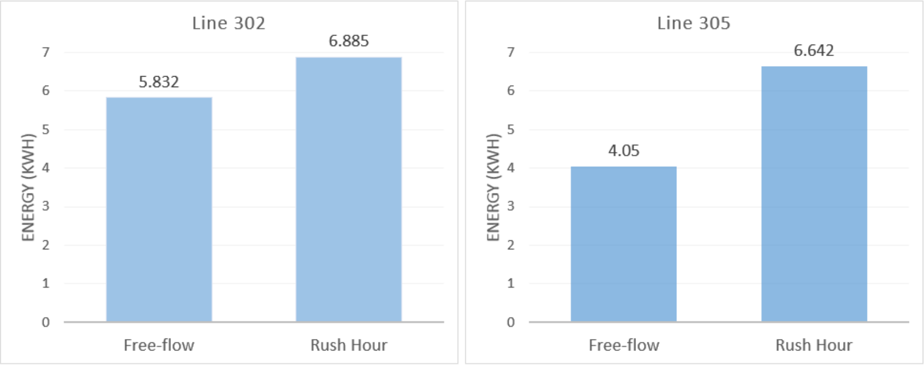


**Fig. 5.** Energy Consumption under different traffic conditions.

As can be seen in Fig. 5 the energy necessary to complete the route 302 with no traffic flow was 5.832 kWh while the energy necessary to complete the same route during the rush hour was 6.885 kWh. The performance of the electric bus in terms of energy is negatively affect when submitted to intense traffic conditions. This impact can also be seen in the results regarding the route 305, where it was necessary 4.05 kWh to complete the driving cycle with no traffic and 6.642 kWh under intense traffic conditions.

Another aspect that can be verified is that the energy consumption impact caused by the intense traffic is less significant for the route 302 than for route 305. In fact, the energy consumption increased 1.053 kWh in the first route which represents an increase of 18% while on the other the increase was of 2.592 kWh, equivalent to a 64%.

A possible explanation to this fact could be related to the difference between routes topology. Line 305 is a route composed of more linear parts than route 302, which allows for the maintenance of a constant speed for larger periods of time when not subject to heavy traffic. Following the same thought, line 302 is characterized for having more curves and that demands more accelerations even when the bus is not immersed in intense traffic conditions.

However, in order to have an accurate analysis, one would have to perform more studies in the influence of route topology on the electric bus performance.

It could be concluded that intense traffic conditions affect negatively the performance of an electric bus, making it spend more energy to perform similar routes and also demanding more power from the bus motor, which implies higher maintenance costs.

## 6 Conclusion

In the last decade, the eco-awareness in our society is growing and our ecological footprint is meant to be as minimal as possible. Therefore, our efforts aim to reduce emissions and optimize electric powered vehicles routes in order to reduce charging time.

This paper describes a HLA-based distributed architecture for electric bus powertrain simulation within a realistic traffic environment. The main intent of this project is to study the impact of rush hour traffic in the amount of energy spent on routes courses by the electric bus powertrain. Another contribution of this work is the description of the process of setting up the simulation scenario based on real traffic data (OD Matrixes) and real traffic flow conditions (Four-Step Model).

According to performed simulations, it could be concluded that heavy traffic conditions affect the performance of an electric bus, demanding more power from the bus and making it spend more energy to perform equal routes, resulting in higher costs. As expected, the most demanding route, with sinuous segments, was also the route that demanded more from the bus, in terms of energy consumption, either without traffic or in heavy traffic conditions.

However, the energy consumption impact caused by the intense traffic is more preponderant in the less demanding route, i.e., the more linear one. A possible explanation is the fact that the less demanding route allows a constant speed for long periods of time when not subject to heavy traffic, and the most sinuous route demands accelerations and decelerations even when the bus is not subjected to intense traffic conditions.

Notwithstanding, one would have to perform more studies in the influence of route topology in the electric bus performance to have a precise analysis. Routes elevations would also affect the bus energy consumption, but due to restrictions of the simulation platform they were not considered in this study. Future work will include a simulation platform with a third axis to take into account the elevation. To allow a more complete performance analysis, more data collection is essential and more parameters must be studied, such as regenerative braking energy and efficiency. This probably will allow a better planning of public transport operations based on electric buses.

## References

1. Arnott, R., Rave, T., Schöb, R.: Alleviating urban traffic congestion. MIT Press Books 1 (2007)
2. Braess, D.: Über ein Paradoxon aus der Verkehrsplanung. Mathematical Methods of Operations Research 12(1), 258–268 (1968), http://dx.doi.org/10.1007/BF01918335

3. Chale-Gongora, H.G., de Weck, O., Doufene, A., Ishimatsu, T., Krob, D.: Planning an itinerary for an electric vehicle. In: 2014 IEEE International Energy Conference (ENERGYCON). pp. 1385–1391. IEEE (may 2014), http://ieeexplore.ieee.org/document/6850603/
4. IEEE Std 1 (ed.): IEEE Standard for Modeling and Simulation (M&S) High Level Architecture (HLA) - Federati Interface Specification. IEEE Computer Society (2010)
5. IEEE std 2 (ed.): IEEE Standard for Modeling and Simulation (M&S) High Level Architecture (HLA) - Framework and Rules. IEEE Computer Society (2010)
6. IEEE std 3 (ed.): IEEE Standard for Modeling and Simulation (M&S) High Level Architecture (HLA) - Object Model Template Specification. IEEE Computer Society (2010)
7. Lei, C.: Cooperative Adaptive Cruise Control model study based on traffic and network simulation. MSc Thesis, Faculty of Electrical Engineering, Mathematics and Computer Science, University of Twente (2011)
8. Macedo, J.L.P.: An Integrated Framework for Multi-paradigm Traffic Simulation. MSc Thesis, Faculty of Engineering, University of Porto (2013)
9. MacKechnie, C.: Electric Buses - An Introduction. Available in http://publictransport.about.com/od/Transit_Vehicles/a/Electric-Buses-An-Introduction.htm, accessed in 18 December 2012
10. Maia, R., Silva, M., Araujo, R., Nunes, U.: Electric vehicle simulator for energy consumption studies in electric mobility systems. In: Integrated and Sustainable Transportation System (FISTS), 2011 IEEE Forum on. pp. 227–232. IEEE (2011)
11. Meyer, M.D., Miller, E.J.: Urban transportation planning: A decision-oriented approach (2001)
12. Perrotta, D., Macedo, J.L., Rossetti, R.J., de Sousa, J.F., Kokkinogenis, Z., Ribeiro, B., Afonso, J.L.: Route Planning for Electric Buses: A Case Study in Oporto. Procedia - Social and Behavioral Sciences 111, 1004–1014 (2014)
13. Perrotta, D., Ribeiro, B., Rossetti, R.J.F., Afonso, J.a.L.: On the potential of regenerative braking of electric buses as a function of their itinerary. Euro Working Group On Transportation (2012)
14. Schrank, D., Eisele, B., Lomax, T.: 2014 URBAN MOBILITY REPORT Powered by INRIX Traffic Data. Tech. rep., Texas A&M Transportation Institute, The Texas A&M University System (2015), http://mobility.tamu.edu
15. Unger, N., Bond, T.C., Wang, J.S., Koch, D.M., Menon, S., Shindell, D.T., Bauer, S.: Attribution of climate forcing to economic sectors. Proceedings of the National Academy of Sciences 107(8), 3382–3387 (2010)
16. Zhang, X., Yang, H.: The optimal cordon-based network congestion pricing problem. Transportation Research Part B: Methodological 38(6), 517–537 (2004)

# Simulation applied to Software Engineering: An Ad-hoc Literature Review

Massanobu Tachikawa[1]

[1] Department of Informatics Engineering
Faculty of Engineering, University of Porto
Rua Roberto Frias, S/N, 4200-465, Porto, Portugal
up201406802@fe.up.pt

**Abstract.** Software process simulation is being used to cope with a diversity of issues from strategic management of software development, to supporting process improvements. The scope of possible software process simulation applications ranges from focused portions of the life cycle to longer term product evolutionary models. This paper reports an ad-hoc literature review of simulation applied to Software Engineering, identifying the questions and issues that simulation could be used to address, what can be usefully simulated, and the approaches and techniques that can be employed. Simulation has been successfully applied in a variety of areas (such as economics, biology, and social sciences), could it then be used to enhance Software Engineering? Papers ranging from systematic literature reviews to practical applications were studied, with the purpose of providing an overview on the simulation applied to Software Engineering, rather than answering defined questions of usual systematic literature reviews.

**Keywords:** Simulation, agent-based simulation, software process modeling, software process, Software Engineering

## 1 Introduction

Simulation is a standard technology in many engineering fields and has been successfully applied in a variety of areas (such as economics, biology, and social sciences). The behavior of complex processes can be analyzed using simulation models and the inherent complexity of software development processes leads to difficulties to anticipate the process behavior and to predict its outcome.

The usual (and costly) way to analyze process behavior is to perform a case study and observe the results, however, usually the results from a particular case study cannot necessarily be generalized to other contexts. Another way of analyzing it is through simulations. Simulation models help to clarify assumptions. They visualize and quantify the assumption about the causes that govern the process behavior and thus support understanding, analysis, and prediction.

We could see the simulation model as virtual laboratories, testing hypotheses about problems, experimenting corrective policies before the implementation in the real system. Observing the experience from applications in other fields than software engineering indicates that real benefits could be gathered by introducing the use of simulation for management decision support.

In practice, process simulation models are frequently used to support project planning and estimation and in this context, simulation is a risk management method, offering not only estimates of cost, but also estimates of cost uncertainty.

The structure of this paper is as follows: Section 2 introduces the background for this research. Section 3 discuss each one of the selected papers. Section 4 provides a discussion of the gathered information from the papers. Section 5 concludes the paper and indicates topics for further work.

## 2 Background

Software Process Simulation (SPS) research area and Software Process Simulation Modeling (SPSM) approach are two key concepts to understand the simulation in the context of Software Engineering and we provide a brief description of these.

Software process simulation (SPS) was introduced to the software engineering community in the 1980s by Abdel-Hamid and Madnick [1] and at that time the SPS technology attracted a great deal of interest from both academics and practitioners, becoming an active research area for managing and improving software development processes.

SPS models were developed to address a variety of software life cycle models including waterfall, incremental, spiral development, and product line development among others.

Continuous simulation (system dynamics, SD), discrete-event simulation (DES) and agent-based simulation (ABS) were the three most applied paradigms in SPS community in the past years [16].

In the past two decades, SPS has demonstrated its value to software practice. Nonetheless, this area is still maturing.

Software Process Simulation Modeling (SPSM) was introduced into the software engineering domain by Abdel-Hamid and Madnick [1] and has been used to address a diversity of issues in Software Engineering.

Kellner et al. suggests that simulation modeling of software processes is particularly desirable when:

- Complexity is beyond human intuition
- There is uncertainty and stochasticity in system behavior
- Risks in process change are very high
- The system has some dynamic behavior
- Decisions made at one point in the system may impact on the process in other aspects

In SPSM, continuous simulation (system dynamics, SD) and discrete-event simulation (DES) are the most commonly used techniques. Moreover, hybrid simulation (SD and DES combined) has also been applied to simulate software processes.

Although simulation modeling has been applied in other scientific and business processes, it is relatively new to Software Engineering practices.

## 3 Selected papers

The papers were selected and organized in the respective sections with the purpose of providing an overview on simulation applied to Software Engineering, rather than answering the defined questions of systematic literature reviews methodology:

- In Theoretical foundations section, we present the paper, which together with [1] introduces the concept of modeling and simulation software process
- In Literature reviews section, papers are listed because we can consider them as valuable condensed information of the area
- Practical approaches to concepts are interesting because they list the applicability and limitations of theoretical concepts
- Theoretical papers list the on going research conducted in the area and the possibilities of integrating the simulation with other paradigms (such as MABS)
- In Book chapters, we can visualize the area within broader concepts such as Empirical Software Engineering and Project Management Software

The process of searching and selecting the papers used the sections listed above as criteria, in addition to the references cited in the papers.

### 3.1 Theoretical foundations

The paper "Software process simulation modeling: Why? What? How?" [8] identifies the questions and issues that simulation can be used to address (`why'), the scope and variables that can be usefully simulated (`what'), and the modeling approaches and techniques that can be most productively employed (`how').

They clustered the many reasons for using simulations of software processes into six categories of purpose:

- Strategic management
- Planning
- Control and operational management
- Process improvement and technology adoption
- Understanding
- Training and learning

Many technical options, considerations and issues remain for the model developer. These are outlined below under the headings of (1) simulation approaches and languages, (2) simulation techniques and (3) data/measurement issues.

The paper concludes by offering some guidance in selecting a simulation modeling approach for practical application, and recommending some issues warranting additional research.

### 3.2 Literature reviews

In [14], more than 200 relevant publications were analyzed in order to find answers a set of research questions and from the analysis the following conclusions could be drawn:

- Categories for classifying software process simulation models as suggested by the authors of a landmark publication in 1999 should be adjusted and refined to better capture the diversity of published models
- Research improving the efficiency of SPSM is gaining importance
- Hybrid process simulation models have attracted interest as a possibility to more realistically capture complex real-world software processes

The primary answers to the first four research questions are:

- Purposes for SPSM (Q1): They can be further detailed as 10 purposes. The cognitive level contains the purposes of 1) understanding, 2) communication, 3) process investigation, 4) training and learning. On the tactical and strategic levels purposes are similar. They are 5) prediction and planning, 6) control and operational management, 7) risk management, 8) process improvement, 9) technology adoption, 10) trade off analysis and optimizing

- Modeling Paradigms (Q2): System dynamics (SD, 49%) and Discrete-event simulation (DES, 31%) were the most widely used techniques in SPSM. Other paradigms included state-based simulation (SBS), qualitative(semi-quantitative) simulation (QSIM), knowledge(rule)-based simulation (KBS), role-playing game, agent-based simulation (ABS), and discrete-time simulation (DTS)
- Simulation Tools (Q3): Due to a large number of SD models developed and published, Vensim is the most popular tool for continuous simulation. Since Extend offers the capability of building both continuous and discrete simulation model, it has been the first choice by the discrete-event and hybrid modelers in SPSM community
- Problem Domains and Model Scope (Q4): 'Project' was the most frequently modeled study scope, particularly for 'generic development'. 'Product evolution' was the next most studied topic

In [15], the author aims to assess how SPSM has evolved during the past 10 years (prior to 2008) in particular whether the reasons for SPSM, the simulation paradigms, tools, problem domains, and model scopes have changed and summarizes, discovers and recommends the 'facts', 'trends', and 'directions' of SPSM research.

Facts: listed above in "Modeling Paradigms (Q2)"

Trends:

- System dynamics and discrete-event simulation form the main stream of SPSM paradigms
- New simulation paradigms continue to be introduced into SPSM research between 2000 & 2005
- Continuous modeling gradually lost its dominant position in SPSM research in comparison with discrete approaches during the decade
- Most of newly introduced paradigms enhanced the research capability at the micro-process level
- In recent years, micro-processes have been attracting more simulation research
- System dynamics and discrete-event simulation are the most common combination for constructing hybrid simulation models
- Hybrid simulation modeling concentrated on vertical integration

Directions:

- More recent modern software development processes need to be further investigated in SPSM research

- More new simulation paradigms need to be experimented and introduced into SPSM community
- More attempts are needed to effectively tackle the uncertainty of software process in practice

A systematic literature review which purpose is to obtain the state of the art of the applications of SPSM in software risk management is presented in [9] with the following conclusions:

- The number of SPSM studies on software risk management is relatively small, but increasing gradually in recent years
- SPSM is mainly applied in risk analysis and risk management planning activities
- Software risks related to requirements, development process and management process are the ones most studied by SPSM
- Discrete-Event Simulation and System Dynamics are two most popular simulation paradigms, while Hybrid simulation methods are more and more widely used
- Extend, iThink and Vensim are the most popular simulation tools in SPSM
- Most of SPSM approaches and models have not been well applied into real-world risk management practices

Update the recent status of SPSM is the objective of [5] and the results of review indicate:

- In terms of the average number of studies found in SE community, there is no significant change compared to the former review period though the number of the studies published in Process Simulation and Modeling (ProSim) / International Conference on Software and System Processes (ICSSP) community declined
- Many Category A studies lack trials of their proposed models on an industry scale problems
- Due to the enhanced capability in capturing the characteristics of complex real-world software processes, hybrid process simulation models have still attracted great interest

### 3.3 Practical approaches

A simulation-based approach to Multi-Agent Systems (MAS) engineering is presented in [10] and discussed its integration within Agent-Oriented Software Engineering (AOSE) methodologies. Integration is first discussed in general by adopting standard method engineering techniques, then detailed by means of a case study—that is, integrating simulation in SODA (explained in [11]).

The process simulation model GENSIM 2.0 is proposed in [6]. GENSIM 2.0 is a generic process simulation tool representing V-model type software development processes. Compared to existing process simulation models in the literature, the novelty of GENSIM 2.0 is twofold:

- Its model structure is customizable to organization-specific processes. This is achieved by using a limited set of generic structures (macro-patterns)
- Its model parameters can be easily calibrated to available empirical data and expert knowledge

The article outlines the structure of GENSIM 2.0, gives examples on how to calibrate the model to available empirical data, and demonstrates its usefulness through two application scenarios The first scenario illustrates how GENSIM 2.0 helps finding effective combinations of verification and validation techniques under given time and effort constraints. The second scenario shows how the simulator supports finding the best combination of alternative verification techniques.

SimSWE, a library of components for modeling and simulation of software engineering processes is proposed in [2]. It consists of a generic, implementation independent description of the components and a reference implementation using the MATLAB / Simulink environment. By providing ready-to-use building blocks for typical functionality, the library should facilitate and ease the use of simulation to analyze software engineering issues.

In [7] is proposed a process simulation model using the System Dynamics modeling technique and used the simulation model in the context of a case study with a software company in Calgary, Canada. With the help of the simulation model, they were able to evaluate the performance of test processes with varying degrees of automation of test activities and help testers choose the most optimal cases. The goal of a earlier study was to investigate how the simulation model can help decision-makers decide whether and to what degree the company should automate their test processes. In this article, the authors present further details of the System Dynamics model, its usage scenarios and examples of simulation experiments independent from a specific company context.

### 3.4 Theoretical papers

Hybrid simulation, the combination of simulation paradigms to address a problem, is becoming more popular as the problems we are presented with become more complex. However, integrating multiple simulation paradigms faces the issues of compatibility, interoperability and synchronization when executing simulation. The objective of [4] is to present the state-of-the-art of this research area, the hybrid mechanism when integrating paradigms, and more importantly provide practical support for the effective adoption of hybrid simulation in SPS context. Based on an extended systematic literature review, this paper presents the state-of-the-art of hybrid software process simulation, and discusses the applicable mechanisms, common combinations and integration strategies when constructing a hybrid process simulation model by using multiple simulation paradigms together.

The authors of [3] propose a model of the software development process that integrates individual-level performance, cognition and artifact quality models in a common simulation framework. In addition, this framework allows the implementation of both Multi-Agent-Based Simulation (MABS) and System Dynamics (SD) simulators using the same basic models. As SD is the dominating approach within SPSM, they are able to make relevant and unique comparisons between it and MABS (reflects the problem domain more realistically than SD).

### 3.5 Book chapters

The chapter [12] aims to raise awareness about the usefulness and importance of simulation in support of software engineering. Simulation is applied in many critical engineering areas and enables one to address issues before they become problems. Simulation – in particular process simulation – is a state of the art technology to analyze process behavior, risks and complex systems with their inherent uncertainties. Simulation provides insights into the designs of development processes and projects before significant time and cost has been invested, and can be of great benefit in support of training.

In chapter [13] , it investigates the reasons for lacking impact of process simulation in the software industry. This is done with the help of an in-depth description of a software process simulation application example. The application example focuses on the effects of various workforce allocation strategies on project performance, expressed in terms of project duration, effort consumption, and product quality. With the help of the application example and based on existing literature, the gap between the current state of the art of software process simulation and the actual state of practice is described and its root-causes are discussed. The chapter concludes with a list of issues that need to be addressed in order to close the gap between the state of the art and the state of practice. Most of the issues relate to the difficulty of demonstrating a positive cost-benefit ratio when applying process simulation as a tool in support of software project management tasks.

## 4 Discussion

Actually, the SPSM is mainly is mainly applied in risk analysis and risk management (relatively small, but increasing gradually in recent years) planning activities, but could potentially be applied in more areas such as strategic management, planning, process improvement and technology adoption, understanding, and training and learning.

Simulation models help to clarify assumptions, visualizing and quantifying the assumption about the causes that govern the process behavior and thus support understanding, analysis, and prediction. Simulation modeling of software processes is particularly desirable when there is uncertainty and stochasticity in system behavior, risks in process change are very high, and the system has some dynamic behavior. In the context of project planning and estimation, simulation is a risk management method, offering not only estimates of cost, but also estimates of cost uncertainty.

More recent modern software development processes and new simulation paradigms need to be investigated, experimented, and introduced in SPSM research, considering that simulation modeling it is relatively new to Software Engineering practices. Specific tools (such as Vensim and Extend) for simulation is another limitation.

## 5 Conclusions and future directions

Answering the question "Simulation could be used to enhance Software Engineering?", simulation modeling it is relatively new to Software Engineering practices and SPS has demonstrated its value to software practice but this area is still maturing. More recent modern software development processes and new simulation paradigms need to be investigated, experimented, and introduced in SPSM research. For instance, most of SPSM approaches and models have not been well applied into real-world risk management practices.

New simulation paradigms continue to be introduced into SPSM research and micro-processes have been attracting more simulation research.

From the practical point of view, the fact that simulations are currently done by tools such as Vensim and Extend could be a limitation for a more massive adoption. The creation of simulation libraries, for more commonly used languages, such as Java or Python, could increase the research community, thus promoting its applicability as a technique.

## References

1. Abdel-Hamid T., Madnick S. E.: Software Project Dynamics: An Integrated Approach. Prentice-Hall, NJ, USA, (1991)

2. Birkhölzer T., Madachy R., Pfahl D., Port D., Beitinger H., Schuster M., Olkov A.:SimSWE - a library of reusable components for software process simulation. Proceedings of the 2010 international conference on New modeling concepts for today's software processes: software process, July 08-09, 2010, Paderborn, Germany.
3. Cherif R., Davidsson P.: Software development process simulation: multi agent-based simulation versus system dynamics. Proceedings of the 10th international conference on Multi-agent-based simulation, May 11-12, 2009, Budapest, Hungary
4. Gao C., Zhang H., Jiang S.: Constructing Hybrid Software Process Simulation Models. Proceedings of 11th International Conference on Software and Systems Process (ICSSP'15), Tallinn, Estonia: ACM, 2015, 157-166.
5. Gao C., Jiang S., Rong G.: Software process simulation modeling: preliminary results from an updated systematic review. Proceedings of the 2014 International Conference on Software and System Process, May 26-28, 2014, Nanjing, China.
6. Garousi V., Khosrovian K., Pfahl D.: A customizable pattern-based software process simulation model: Design, calibration and application. Software Process Improvement and Practice, vol. 14, n. 3 (2009) 165–180
7. Garousi V., Pfahl D.: When to automate software testing? A decision-support approach based on process simulation. Journal of Software: Evolution and Process 28(4) (2016) 272-285
8. Kellner, M.I., Madachy, R.J., and Raffo, D.M.: Software process simulation modeling: Why? What? How? Journal of Systems and Software (1999). 46(2-3): 91-105.
9. Liu, D., Wang, Q., Xiao, J.: The Role of Software Process Simulation Modeling in Software Risk Management - a Systematic Review. Proceedings of the 3rd International Symposium on Empirical Software Engineering and Measurement (ESEM'09), Lask Buena Vista, FL, IEEE Computer Society (Oct. 2009) 302–311
10. Molesini A., Casadei M., Omicini A., Viroli M.: Simulation in Agent-Oriented Software Engineering: The SODA case study. Science of Computer Programming, v.78 n.6, (June 2013) 705-714
11. Molesini A., Omicini A., Ricci A., Denti E.: Zooming multi-agent systems. International Workshop on Agent-Oriented Software Engineering (2005) 81-93
12. Müller M., Pfahl D.: Simulation Methods. F. Shull, J. Singer, D. Sjøberg (Eds.), Guide to Advanced Empirical Software Engineering, Springer, 2008
13. Pfahl D.: Process Simulation: A Tool for Software Project Managers? Ruhe G., Wohlin C. (Eds.), Software Project Management in a Changing World, Springer, 2014
14. Zhang, H., Kitchenham, B., Pfahl, D.: Reflections on 10 years of software process simulation modelling: A systematic review. Proceedings of International Conference on Software Process (ICSP'08). Volume LNCS 5007., Leipzig, Germany, Springer-Verlag (May 2008) 345–365
15. Zhang, H., Kitchenham, B., Pfahl, D.: Software process simulation modeling: Facts, trends and directions. Software Engineering Conference, 2008. APSEC'08. 15th Asia-Pacific. IEEE (2008) 59-66.
16. Zhang H., Raffo D., Birkholtzer T., Houston D., Madachy R., Münch J., Sutton S. M. Jr:: Software process simulation—at a crossroads? J. Softw. Evol. and Proc., 26, (2014) 923–928

# SESSION 3

## INTELLIGENT SYSTEMS

**Epidemic disease: analyzing twitter message behavior with multi-agent system**
*Denis Ryoji Ogura*

**The Banking Internal Audit Framework**
*Gudilak Constantino*

**Interactive Solution for Multimedia Content Evaluation**
*Tiago Soares Costa*

# Epidemic disease: analyzing twitter message behavior with multi-agent system

Denis Ryoji Ogura

School of engineering, University of Porto, Porto, Portugal
denis.ogura@fe.up.pt

**Abstract.** There are several works using social networks as a tool to collect data and analyzes the user activities to achieve goals such as: marketing inputs, sentiments of a product or service, evaluation, etc. In this directions there are some studies reading social networks to finding disease outbreaks in advanced. This works aims to provides a methodology that gather tweets with disease keywords, and pass through a method to analyze each tweet, classifying as disease or not disease. In the methodology were defined to use multi-agent system (MAS) to process in distributed application approach and possibly the process to use agent to execute independently each other. For the analyzer process, were collected ~83 thousands tweets for gathering the disease keywords inputs, which is fair enough to find a disease outbreak in a certain country. The classification process we defined to use supervised learning algorithm (K-nn) which provided a good method to categorize the tweet message as disease or not disease. Finally we reported the output of the methodologies finding cancer (67.47%) and flu (9.72%) keywords as a starting point as a possibly disease outbreak.

**Keywords:** Multi-agent system, data mining, text mining, twitter.

## 1 Introduction

The social network has become a mechanism for sharing information, photos, status, location, and connect people virtually. This new paradigm has been explored by companies to collect certain information and define company strategies based on information shared. Based on collecting behaviors, raise possibilities to analyze user activities [7] and define methodologies, models, architectures, and systems for gathering solutions based on data collected.

This study aim to provide a mechanism to predict [3,8] if a Twitter text message has relations to a disease or not, and further analyze conclude if a country or region is having starting evidence of an epidemic disease outbreak. Of course this method depends on user post text related to sickness or proving symptoms with the keyword defined to collect.

This method can be useful for health government body to get in advanced possibly epidemic disease outbreak and take actions in advanced on finding vaccine, request drugs budget, or even purchase drugs based on the trend of the twitter. Normally, the epidemic it is announced after patient goes to medical assistance in the hospital, or emergence in health centers. After some times, health government body raise an health alarm reacting to those patient requested assistance. Since this work intention is to provide a more proactive method, it is positive to provide a methodology to take action in advanced and achieve the goal to anticipate the disease outbreak and make the health care system more sensitive.

Although this method depends on user activities, if users don't post anything, this methodology does not work at all. The main advantage of this methodology, it can be integrated with others social networks, such as: Facebook, Instagram, Snapshat [12]. The input data is well documented and easy integrate with CSV extension file. For applying new social network in the methodology just need to export data with the data export model applied in this method.

To make this work more reliable, we included MAS (Multi Agent System) framework approach for making the collect method be separated in agents to work in parallel (threads) and distributed. With this method, the project gain flexibility, scalability, and more efficiency since can be distributed to run in different computers. For the classification process for a unknown message is a disease related or not, we used RapidMiner [14, 16] tool to process with text mining for classifying using K-nn Cosine algorithm.

This work is organized as follows: Chapter 2 is regarding the related work of this study, Chapter 3 provides the methodology of the work and concept, Chapter 4 is the analyzes of the results collected data from Twitter and Chapter 5 is future work and conclusions.

## 2 RELATED WORK

There are several works that has been carried out in the analysis data of the user activities in social network, which enabled research further in this area to gather information for institutional purpose, marketing, business, customer feedback, customer satisfaction, etc. Based on users activities in social networks, enable the collection and analysis using different approach and tools. Del Val et al. [7] provided a mechanism of gathering data from users activities in social network using Multi-agent systems. The authors performed an in depth analysis of the evolution of social behavior at different levels in online events and calculated metrics based on network theory. Ahmad et al. [2] provide a text mining method which classify large amount of data. They presented a distributed document classification technique using Multi-agent technology. For classification process, they used Naive Bays algorithm for supervised learning method. Tripathi et al. [17] worked on classifying tweets from Twitter and make a analysis of the sentiment of it. The authors applied data mining techniques that applied text mining to make a classification process with the attributes happy, sad, and neutral sentiments for a particular tweet. They compared two different classifier K-nn (K-Nearest Neighbor) and Naive Bayes. Romano et al. [15] made a study to use social networks for a early detection of public health events, for epidemic outbreaks. They collected a massive dataset of tweets, and then extracted the relevant information of different outbreaks from different countries in 2011. The challenge was define a strategy to considerer messages due to a variability in the temporal dynamics of twitter messages from different diseases, and the identification of a suite source of message to made the classification possible, and to define method suitable for the assessment of the analysis algorithms. Quwaider et al. [13] provided a study of disease are associated with Ebola, Malaria, and HIV/AIDS that are posted in Facebook. The authors created a benchmark dataset for disease, called BDDC (Benchmark Dataset for Disease Classification). The benchmark is well documented and compatible with recognized text mining tools and can be comparable with others studies. This work, differently of the works above, is intended to provide a mechanism of collect tweets with certain disease keywords, and then pass through a text mining process for classifying using supervised learning algorithms (K-nn) for indicating if a tweet message is disease related or not. Based on this classification, and the location of the user, it is possible to conclude if a region is a possibly starting issue of a epidemic disease outbreak. Following sections we are providing information about the methodology and strategy for classification process.

## 3 METHODOLOGY

The methodology of this work is intend to provide all the component required to predict a disease epidemic outbreak. We developed the collection, and analyze data using MAS framework and take advantage of processing distributed and parallel approach. Each phase run in an agent and processed via behavior method. The multi-agent system can distribute the processing of data on computer nodes, it is more reliable to process large amount of data that has been collected by Twitter4J API [6,9] which is the API to connect to Twitter database and collect tweets in twitter streaming database.

Some issues were found in Twitter API:

- The location of the User is written by the user itself. It was found text messages, city, county, village, country, and null. Those information is not precise on where the tweet message was created from. The methodology tries to search in the location field if find any key that can redirect to the country name. The solution will be explain in the Chapter 4.
- We tested the latitude and longitude field, to be precise on where the message was created from. However, we realized that the user must turn on the GPS feature to provide the exact latitude and longitude position in the globe. The amount of messages we collected with GPS location feature is not enough amount of message for applying in this work, then we decided to not use this field.
- The Twitter history database has limit transaction for searching in it. For gathering the database access unrestricted must have permission from Twitter APP. However, the timeframe of the production of this study, made unfeasible to request

the special access for research purpose. Then we decided to move forward with streaming message only, since there is no limit to listening and store messages are coming from the streaming.

Initially, the collection of the tweet was defined by disease keywords only, however, the twitter streaming returned an amount of messages in several languages. To avoid messages that is not understandable, as well as for analysis process, we filtered messages that are written in English language only. In the programming used the attribute as followed: `twitter4j.Status.getLang().equals("en")`. Also, for location with no value (NULL), it's disconsider since the location is important for this work. Following Table 1 the keywords we are using to collect tweet messages.

**Table 1**: Disease keywords for collection process

| | | |
|---|---|---|
| Alzheimer | Anthrax | Asthma |
| Cholera | Dengue | Diabetes |
| Ebola | Flu | Gonorrhea |
| Influenza | Leptospirosis | Leukemia |
| Meningitis | Migraine | Multiple sclerosis |
| Rubella | Strep throat | Swine |
| Ulcers | Yellow fever | |

Following some examples of tweet messages that are considerer as disease and not disease related:

**Table 2**: Tweet message not related to disease

| **Non disease tweets** |
|---|
| Non stop horror - thoughts are with everyone around the globe suffering from the **cancer** of terrorism and persecution |
| I may get some form of **heart disease** from all this salt these liberals are giving me today.#ElectoralCollege #electors |
| When people come at you sideways not realizing you're a..Libra???**Cancer**???Pisces???Aries???Scorpio???Capricorn??? |
| Ty tryna make me go into an **asthma attack** like nigga what's wrong with you??? |
| Just got a notification that this is now on Netflix. This just gave me **cancer** |

In the Table 2 shows the tweet messages mention the keyword of the disease, however it's a metaphor or another meaning, which of course do not represent the disease itself. The user just wanted to represent the felling using disease terms, or the example of the row 3, which is the cancer zodiac sign that of course is not the cancer disease reference. So those examples in Table 2 are clearly tweets messages that is not related to a disease at all.

**Table 3**: Tweet messages related to disease

| **Disease tweets examples** |
|---|
| I think I'm having **asthma attack** oh well |
| it's not even for my cold either. It's for my **asthma**, etc. |
| my moms crying bc my dad has **diabetes** and is going to die. first of all he is borderline diabetic and isnt going to die? its not severe? |
| When u have **bronchitis** and the flu and have no insurance AND can't even afford any medicine or soup or anything to help hahahaha wow |
| South Korea very concerned over bird **flu** outbreak, aftermath: vice finance minister |

In Table 3 provide some examples of tweet messages, advertising, informative, thoughts or text related to a disease. With this manual initial samples, would be possible to classify tweets messages are coming from the collect process and use text mining (K-nn algorithm) to classify if it is a disease related or not. The process of classification method will be shown in another investigation with further information, and comparison on using different algorithms of classification process and with the best method of classify tweet message related or not related to a disease.

### 3.1 MAS framework

In this section lets explore the MAS (Multi-agent System) framework for the collection of tweets, analysis, and finally classify it under disease or not disease taxonomy. The main purpose of using MAS is to provide a distributed processing methodology and make the project more flexible, scalable and distributed processing. With that said, we can process each agent system under different machines, and increase the number of computational resources for processing a certain amount of data. This project uses MAS framework with Jade platform [1,5,14], and were defined agents rules to process behaviors methods.

Jade API is an agent language method that is written in Java language with all necessaries libraries and resources to create agents, containers, behaviors, and apply functions to be executed by agent via behaviors methods. Jade API has shown a good strategy of applying agent and distribute the processing in multiples tasks, threads, so could use a distributed infrastructure to process this entire project.

For this work, were defined four agent: collector, analyzer, message dispatcher, and decision maker. The collector agent is responsible to gathering tweet messages that contains the disease keywords Table 1, using the Twitter4J API to reach the Twitter streaming database. The process will collect 5000 tweets per cycle, and generate a file to be processed in the analyzer agent afterwards. The message dispatcher agent has the role of receive this file from twitter agent and dispatch the file to analyzer agent. The analyzer agent will be trigger by dispatcher message agent to startup the analyzer agent to classify the 5000 tweets messages. The decision maker agent is intended to execute a training set with more 50 samples of tweet related to a disease or not. If the accuracy improves, apply it in the classification process. With that applied, the analyzer process will classify a unknown tweet message taxonomy more accurate. However, if loose accuracy, we defined that 6% or less it is good percentage loss to apply anyway, since with more sample, the false positive might increase, but do not affect the classification process. Indeed, applying this 50 samples of disease and not disease will be applied to classification process as well. Otherwise, disregards those samples.

Following we provide AOSE [4,11] (Agent Oriented Software Engineering) diagrams and description to document the project with MAS architectures. All containers and agents are created at main program, but all the behaviors are defined in each agent class setup() method.

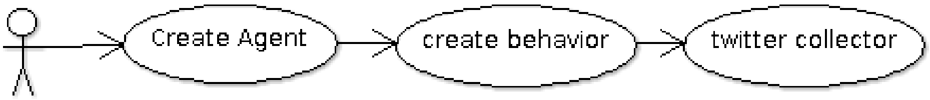


Figure 1: Use case of twitter collector agent

This Figure 1 shows the use case of the twitter agent collector, providing the process of instantiate the agent and trigger the behavior to starting collecting tweet message from twitter streaming database.

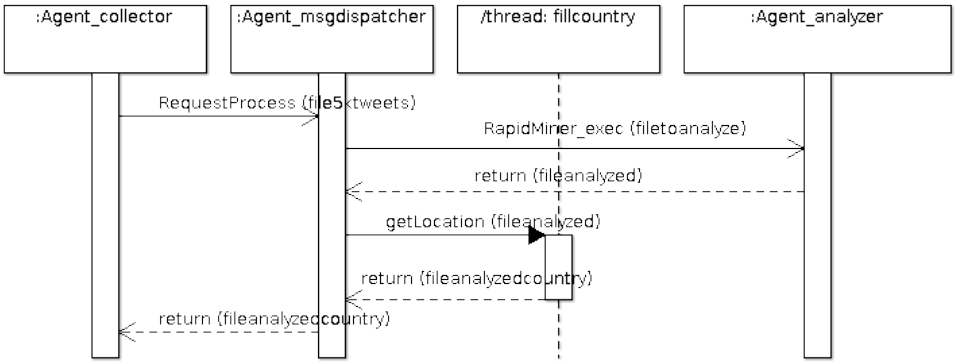

:Agent_collector
:Agent_msgdispatcher
/thread: fillcountry
:Agent_analyzer
RequestProcess (file5ktweets)
RapidMiner_exec (filetoanalyze)
return (fileanalyzed)
getLocation (fileanalyzed)
return (fileanalyzedcountry)
return (fileanalyzedcountry)

Figure 2: Sequence diagram of the agent communication order.

The Figure 2 provides the sequence diagram of the process for analyzing the 5000 tweets messages were collected by the Agent_collector. Soon as the Agent_collector collects 5000 tweets, send a message to Agent_msgdispatcher requesting to process the file file5ktweets. The agent_msgdispatcher will send a message requesting to Agent_analyzer agent to call RapidMiner_exec function passing the file5ktweets as parameter. This function will classify each tweet message using supervised learning algorithms (K-nn cosine) to categorize each tweet message as disease or not disease. The result of RapidMiner_exec process will create another file (fileclassified) which will return to agent Agent_msgdispatcher. Next step of the process will create a thread to execute the method in parallel of the agent process. This thread will execute the method getLocationFiles() which will fill in the country identified based on the user location field. Finally, the process return with the file and all tweets messages categorized as disease or not disease, with the country found.

Following all the roles and responsibilities for each agent:

Table 4: Agent Roles and responsibilities

| Agent | Roles and responsibilities |
|---|---|
| Agent_collector | • Add behavior to execute twitter collector API<br>• Manage itself to split 5000 tweets per cycle<br>• Filter tweets by disease keywords Table 1 |
| Agent_msgdispatcher | • Manage new files within 5000 tweets messages<br>• Dispatch the file to analyzer agent<br>• Wait for the reply back from the analyzer agent within the output of the classification process<br>• Validate the confidence of the classification process and update the sample of groups disease and not disease.<br>• Identify a possibly focus of epidemic disease<br>• Identify Country based on User Location |
| Agent_decisionmaker | • Gather disease and not disease samples and include in the classification process. If the increase the accuracy or reduce less or equal than 6%, will apply this new example set. Otherwise disregards. |
| Agent_analyzer | • Receive a file with tweets messages unclassified<br>• Execute RapidMiner process to classify the tweet message using supervisor learning algorithms.<br>• Return the output of the classified tweet messages<br>• Verify the percentage confidence of the training set<br>• Generate disease and not disease tweets samples to increase the confidence of the training set |

In the next section we will explore the experiments, results and analyzes of tweets messages collected using the methodology provided in this work.

## 4 EXPERIMENTS AND DATA ANALYZES

The aim of this section is to describe scenarios of the experiment execution, data gathering, timeframe, and data analyzes of collected data to validate the confidence, as well as identify a possibly starting focus of epidemic disease for specific location. Soon as execute the Java project with JADE API, start the Jade GUI manager that will contains all the managed containers, agents, behaviors, send and receive messages, ACL message protocol, and all resources required to a agent execute properly. The main program will create all necessaries resources in Jade manager GUI, which is in different containers for each agent, to be easy to implement in a real distributed infrastructure. After all resources are created, start the sequence logic as detailed at Figure 2.

The RapidMiner process has a percentage confidence of the training set that is important to validate how confident is the training set phase to classifying tweets into groups. As defined in the beginning of this project, were manual selected 300 tweets related to disease and 300 tweets as not disease. The number 300 tweets is just a starting point. The training set represent 77% of the accuracy rate and 22% of classification error, which is good percentage as starting point.

Regarding the Agent_DecisionMaker agent it has a process to execute the method TickerBehavior() passing 600000 miliseconds (10 minutes) as parameter to execute 10 in 10 minutes. This method gather 50 samples of disease and not disease classified in the analyzer process. Then, these samples are included in the training set and execute it using K-nn algorithm and validate if improves the accuracy. If increase the accuracy percentage or do not loose more than 6% will apply these 100 tweets message into classification process. Since it is increasing the number of example set, may decrease the accuracy since there more samples to get false positive, true positive, false negative and true negative. The main purpose of this method is to have more example set to improve the training set to be more accurate as possible. So the Agent_DecisionMaker agent has the ability to decide if apply 50 more sample of disease or not disease tweets messages into the classification process.

After pass through all the steps above, we found some challenges in regards the Twitter user location, since the user is not obligated to inform where is the actual location. Some users add phrases, misspell the city name, and some cases inform city, state and country. For this study is required the country for the analyze phase, indeed used YQL [10] (Yahoo Query Language) searching into `geo.places` data table which contains all cities, states, countries names. Following a example of query and results:

Table 5: Examples of YQL query

| YQL query statement | Results |
|---|---|
| select * from geo.places where (placetype='Country' OR placetype='Town' OR placetype='State') and text='Toronto, Canada' \| truncate(count=1)" | "country": { "code": "CA", "type": "Country", "woeid": "23424775","content": "Canada"} |
| select * from geo.places where (placetype='Country' OR placetype='Town' OR placetype='State') and text='Anywhere in the globe' \| truncate(count=1) | "results": null |

However, this method brings some threats, if the user set any text that contains a city, state or country in the location field, the YQL will return as shown in the Table 6. As we known the method is not precise, but we decided to move forward with this method, because works for most of the location and figure for most of all tweets messages. We did a manual investigation of those tweets with message in location field and it is represent less than 0.1% of the entire data set.

Following some query examples that send a text message, and the YQL returns a country since there is a word that represent and city, county, state or country name:

Table 6: YQL queries with anything in location bring a country

| YQL query statement | Results |
|---|---|
| select * from geo.places where (placetype='Country' OR placetype='Town' OR placetype='State') and text='Over the Hills and far away' \| truncate(count=1) | "name": "The Hills","country": {"code": "US","type":"Country","woeid": "23424977","content": "United States" |
| select * from geo.places where (placetype='Country' OR placetype='Town' OR placetype='State') and text='Whatever British' \| truncate(count=1) | "name": "British","country": { "code":"GB","type":"Country","woeid" : "23424975","content":"United Kingdom" |

Moreover, there is a process to fill the country in the tweet record in file, and then go through a filter process which exclude those tweets are not related to a disease, and tweets that not contains any disease keyword in the text message (Twitter4J sometimes does not respect the query criteria, and brings messages that are not related to any keyword anyway), but also was classified as disease by the analyzer agent. Those message are listed as `yesdisease`, and the disease keyword can't be found in the text message, is a good example of the classification process must be updated in the `nondisease` group in the training set. Afterwards, has left ~83 thousands tweets with the criteria of `yesdisease`, and there is at least one disease keyword in the tweet message. With those text message were analyzed and provided in the following charts.

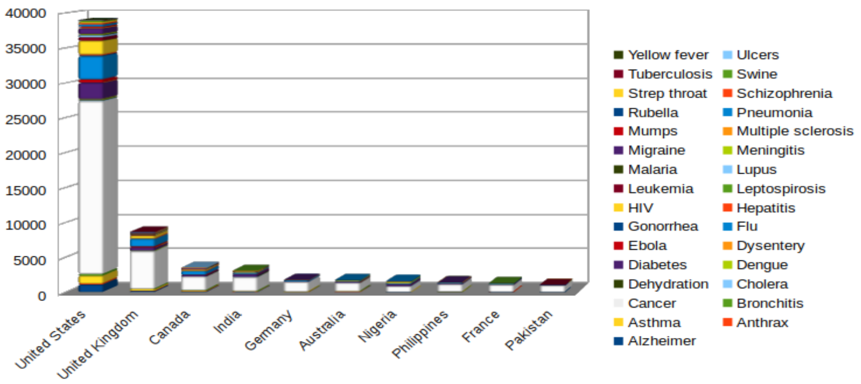


Figure 3: Top 10 of Countries vs Disease

The Figure 3 has a chart with the top 10 countries with disease encountered in the text message. We analyzed the results for all countries and realized that cancer disease is the most keyword has found for all countries. To make the chart more readable, we filtered by top 10 countries. Futhermore, we looked into results and found the second disease keywords is flu, which make sense since there are countries in winter season at this moment. To ensure that cancer and flu are the two main disease keywords has found in ~83,000 tweets message, we made another chart to illustrate the findings.

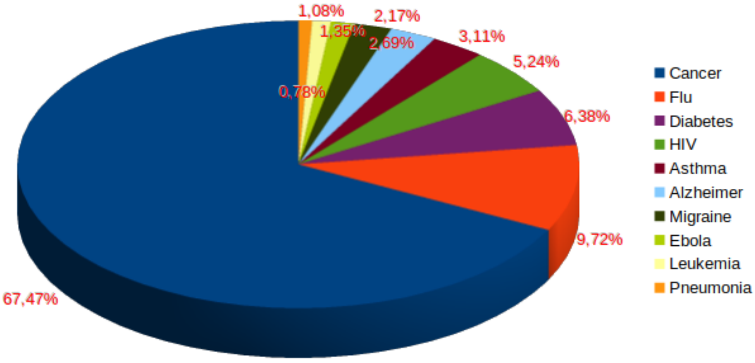


Figure 4: Top 10 disease

The Figure 4 provides another view of the disease summarizing per disease keyword all the tweets messages. One more time, Cancer were found in 67% of the tweets messages, and Flu with 9%. So, in the entire world, most of the tweets are using keywords with cancer and flu.

## 5 CONCLUSION AND FUTURE WORK

This study has shown the twitter social network is a good strategy to analyze tweets messages and evaluate if a country is passing through a moment with high amount of tweets related to a disease. With this methodology, health government body could gathering these disease input from the twitter social network and act in advance. Nowadays, persons use social network to share status, advertising, information about a disease, and it is good strategy to use those information and take in advanced the disease outbreaks.

The goal of this study is to provide a methodology of gathering data from Twitter streaming database using MAS approach to collect data. Afterwards, pass through a classification process to categorize the tweet message as disease and not disease related. The analyze provided has shown that cancer (67%) is the most disease found in the all countries. Which means that all health government body of all country may investigate further on cancer disease, since is most disease commented by the Twitter users. In regards the second disease, flu (9%) is very common disease outbreak in winter season, so the health government body may be prepared for flu disease every year. One last observation, United state and United Kingdom are the highest amount of tweets because we are only analyzing tweets in English. Maybe if we make another investigation with Spanish, Portuguese, might bring more countries to compare with.

This study has worked based on the methodology of gathering and classifying each tweet message collected in regards to disease keywords. Using MAS approach has shown good strategy to process the project in parallel and distributed method, which enabled scalability, efficiency and flexibility in the computer environment. The analyze process has shown an accuracy of 77% and 22% of the classification error as starting point, which may explain some tweets were been classified as disease but in fact is not a disease related. This is  a evidence that this process must be investigate further as future work. Also, it is required to execute this project in a large scale computer environment to evaluate the performance gain of using multiples machines. For classification process would be interesting to execute with others supervised or unsupervised learning algorithms such as Naive Bayes, SVM, decision tree, clustering algorithms, etc. In addiction, uses different disease keywords and analyze the behavior of twitter users in compare to cancer and flu. Another possibility is analyze only user tweet message (excluding advertising, chain and information message) and then pass through the process to validate if cancer and flu remain the main disease outbreaks.

## REFERENCES

1. Abdulrahman R, Neagu D, Holton DRW (2011) Multi Agent System for Historical Information Retrieval from Online Social Networks. 54–63.
2. Ahmad R, Ali S, Kim DH (2012) A Multi-Agent system for documents classification. ICOSST 2012 - 2012 Int Conf Open Source Syst Technol Proc 28–32. doi: 10.1109/ICOSST.2012.6472823
3. AL-Mutairi HM, Khan MB (2014) Predicting the Popularity of Trending Articles in the Arabic Wikipedia using Data Mining Techniques. Proc 6th Int Conf Manag Emergent Digit Ecosyst - MEDES '14 204–205. doi: 10.1145/2668260.2668304
4. Castro A, Oliveira E (2008) The rationale behind the development of an airline operations control centre using Gaia-based methodology. Int J Agent-Oriented Softw Eng 2:350. doi: 10.1504/IJAOSE.2008.019423
5. Castro A, Oliveira E (2008) The rationale behind the development of an airline operations control centre using Gaia-based methodology. Int J Agent-Oriented Softw Eng. doi: 10.1504/IJAOSE.2008.019423
6. Chalmeta-rosalen R (2010) Industrial Management & Data Systems. Doi: 10.1108/02635571011020331
7. Del Val E, Martínez · C, Botti · V, Herrero AB, Martínez C, Botti V (2016) Analyzing users' activity in online social networks over time through a multi-agent framework. Soft Comput 20:4331–4345. doi: 10.1007/s00500-016-2301-0

8. Fu Q, Hu C, Xu W, He X, Zhang T (2014) Detect and analyze flu outlier events via social network. Lect Notes Comput Sci (including Subser Lect Notes Artif Intell Lect Notes Bioinformatics) 8710 LNCS:136–147. Doi: 10.1007/978-3-319-11119-3_13
9. Gemci F, Peker KA (2013) Extracting Turkish tweet topics using LDA. Electr Electron Eng (ELECO), 2013 8th Int Conf 531–534. doi: 10.1109/ELECO.2013.6713899
10. Iturrioz J, Azpeitia I, Díaz O. "Cross Publishing 2.0: Letting Users Define Their Sharing Practices on Top of YQL." International Conference on Web Engineering. Springer International Publishing, 2014.
11. Nikraz M, Caire G, Bahri P (2006) A methodology for the analysis and design of multi-agent systems using JADE. Syst Sci 41.
12. Phua J, Jin SV, Kim J (Jay) (2017) Gratifications of using Facebook, Twitter, Instagram, or Snapchat to follow brands: The moderating effect of social comparison, trust, tie strength, and network homophily on brand identification, brand engagement, brand commitment, and membership intention. Telemat Informatics 34:412–424. doi: 10.1016/j.tele.2016.06.004
13. Quwaider M, Alfaqeeh M (2016) Social Networks Benchmark Dataset for Diseases Classification. 2016 IEEE 4th Int Conf Futur Internet Things Cloud Work 234–239. doi: 10.1109/W-FiCloud.2016.56
14. Ramadan EF, Shalaby M, Elfakhrany E (2016) Cooperation among Independent Multi-Agents in A Reliable Data Mining System. 48–52.
15. Romano S, Martino S Di, Kanhabua N, Mazzeo A, Nejdl W (2016) Challenges in detecting epidemic outbreaks from social networks. Proc - IEEE 30th Int Conf Adv Inf Netw Appl Work WAINA 2016 69–74. doi: 10.1109/WAINA.2016.111
16. Tripathi P, Vishwakarma SK, Lala A (2015) Sentiment Analysis of English Tweets Using RapidMiner. 2015 Int Conf Comput Intell Commun Networks 668–672. doi: 10.1109/CICN.2015.137
17. Uzunkaya C, Ensari T, Kavurucu Y (2015) Hadoop Ecosystem and Its Analysis on Tweets. Procedia - Soc Behav Sci 195:1890–1897. doi: 10.1016/j.sbspro.2015.06.429

# The Banking Internal Audit Framework

Gudilak Constantino

FEUP, Rua Roberto Frias, Porto, Portugal
Gudilak@live.com.pt

**Abstract.** Today's organizations are producing more information in electronic format than never before. This paradigm, on which auditors have to deals with electronic evidences, raises the necessity of organizations to perform auditing in (more) maturity way. Thus, enable audit professional with systematic methodology, measures and metrics, complementary supported by a technological tool is the overall objective of this work. Such holistic perspective face toward to plan, carry out and guide audit actions within an improvement continuity perspective. Adopting in a preliminary stage a set of case studies as research approach allowed to elicit the requirements that inform the presented proposal. As so, this work yields a framework to guide a systematization of audit life cycle. The principal contributions are: 1) The Conceptual framework to accompaniment and track the life cycle of auditing, 2) Its integration in the multidimensional data modeling paradigm, which together 3) allow the optimization of work and a guidance for development of a tool that supports the proposed approach. A questionnaire was submitted to auditing experts, allowing to assess the created model. The results of the conducted assessment revealed a satisfactory perspective of experts toward the present work contribution.

**Keywords.** Bank Audit; Audit Life-Cycle; Audit Route; Business Intelligence; Multidimensional Data Modeling.

## 1 Introduction

Nowadays, organizations are facing very competitive environments, thus stakeholders require guarantees of proper performance, as so, auditing processes become powerful instruments to determine operational and financial quality of corporate organizations [1], [2]. Therefore, and considering that the scope of this work relies on bank institutions auditing (particularly on the Angolan market), [3] provides the recommendation of auditing endeavors in bank institutions, as very own banks' life's assurance. For an effectiveness and efficient audit function, bank´s audit teams should consider on their approach an audit methodology and the data analytics-enabled auditing perspective [4]. Data analytics-enable auditing is a way to automate audit process, enabling audits that are more focused on risk-based data gathering and a more efficient analysis of larger population [5]. Despite the benefits of data analytics-enabled auditing, audit teams should, simultaneously, focus on audit sound audit methodology, in other words, auditors must pay adopt and/or develop the methodology to provide guidance for auditing processes, that is, define criteria on planning, execution, and reporting on a continuous and integrative basis [4].

This research consist in the proposal of a framework for methodological guidance of auditing tasks within a full spectrum: from underlying conceptual proposal to a Computer Audit Assisted Tool aids Development (CAATD). Thus, from conceptual point of view, is defined criteria to systematize an audit process, that is, define specific criteria (concept, metric and indicator) that allow track/monitor different phases of an auditing process, in this paper defined as audit life cycle, and principles that allow track different kinds of audits, which in the present proposal is formulated as the audit route concept. The present proposal is further furnished with its integration within the multi-dimensional data modeling orientation, underlying Data Warehouses (DW) development, which are a hardly neglected data source to feed audit endeavors. Such combination: conceptual framework for audit processes and systematic integration with DW, yields the specification for CAATD aids through the definition of principles and instruments to coupling the conceptual model with the multi-dimensional modeling concepts. Although a proof of concept prototype is briefly presented in the paper conclusion, fine grained data modeling process and implementation issues are not object of this paper.

### 1.1 Methodology

Initially, one conducted a study of the Angolan financial market, which has culminated with its characterization. From structural point of view, it was found (in the analysis period - 2012) that, the Angolan bank system was composed by 23 banks, which are classified regarding the nature[1] and dimension[2] of them. It was noted that, there was an increase in the main operational indicators (deposits and credits) and a little bit decrease (in order of 9%) in the financial (profitability) indicators.

The characterization of Angolan financial market, allow us to establish and select 3 "cases" considered informative for current research purposes. Interviews, qualitative questionnaires and documental analysis constituted the instruments to collect data, for each "case". The information was analyzed accordingly the following dimensions: **characterization of institution** that is, regarding it by major operational nature, dimension and market share; **audit direction structure** - focusing on organic structure of the audit direction, describing its constitution areas, activities and skills; **audit types and objects -** namely, its scope, goal, and driven activities; and **audits plans and respective management -** exposes the way of how audits are planned, executed, implemented, and managed.

Based on the conclusions taken from these cases' dimensions and existing best practices the conceptual model was built. To illustrate the concepts proposed in the model, one selected "term account" product, which is a basic financial operation through which banks raise resources. Thus, to validate the conceptual model a struc-

[1] The nature of banks is related to the form of legal representation, it can be public nature, mixed, private national, and foreign bank branch

[2] The bank dimension is defined according to their representation in the aggregated actives of the system.

tured questionnaire was developed and administrated to auditing experts/professionals.

This paper is organized in six sections. The current section accomplish the introduction to the subject and scope, where is shortly addressed the research goal and the methodology adopted. The second section brings up another dimension of the work foundations which is the existing related work collected from a literature review. Section three, proposed framework, come up with succinct description of the project coverage and then deepens on presenting the framework concepts. The section four describe concepts and instrument to integrate the proposed conceptual model with multidimensional data modelling concepts. In section five it is outlined the evaluation and validation of the model and discussed the results. Finally, in section six we draw the conclusions and point future work directions.

## 2 Related Work

As response to the current organization environment companies are rapidly moving toward more automated environments/systems regarding auditing assistance/support [6], [7]. A classic approach is presented by [8], that suggest a shift of paradigm adopting Continuous Audit (CA) and Continuous Monitoring (CM) concepts to efficiently and effectively conduct and evaluate auditing endeavors. On one-hand, CA is defined as an auditing methodology for report simultaneously, or shortly after, a relevant event occurrence [9], [10]. On other-hand, CM is "*a process implemented by management to ensure that business is operating effectively"* [9]. The study conduct by [7] reinforce the idea of integrating CA/CM on a systematic auditing methodology. This study affirm that the adoption of CA/CM methodology enhance audit process assurance and reduce external auditing cost. In the same way, [11] emanate that the CA/CM methodology improve the organizational oversight status and enhance the reliance of external auditors. The CA/CM methodology is also contemplated in the present work, that is, based on this methodology the proposed framework enable auditor team to track and/or keep up with audit life-cycle and audit route evolution.

Although CA/CM methodology enhance the audit process, one can acknowledge the value of a step forward considering the suggestion to integrate the audit process with a Computer Assisted Audit Tool (CAAT[3]) [12]. Another acknowledgement is given by [13], that consider the use of CAAT as optimized level in the auditing information system maturity level. Accordingly [14], [15], point out two reasons for using CAAT: identify correct data to be audited and ensure the data analysis is relevant and complete. Identify, polish and structure the correct data is a fundamental to auditing reliability. Therefore [16], [17], for example, suggest the build of a Data Warehouse (DW) to aids this purpose.

Another objective of a CAAT is to provide relevant information to the auditor and other decision makers in a customizable/suited manner. Such approach is touched in [17], where the authors suggests a CAAT architecture based on tree tier, namely: DW,

[3] CAAT can be defined as use of computers resources to achieve and/or automate audit process.

OLAP Sever, and Visualization. A similar idea is suggested in [4]. The author fall back on multidimensional data modeling to provide an illustration on how to implement sustainable data analytics-enabled auditing. The idea of use CAAT three tier architecture and multidimensional data modeling was also adopted in this research for illustration propose. The positive impact in audit activities of the coupling between CAAT and CA/CM has been reported in [6]. This context become the work foundation for the framework proposed in this paper.

## 3 Proposed Framework

An audit process life-cycle, is characterized by the whole activities developed in the different auditing phases: planning, execution, and corrective actions evaluation. The audit route concept, consists in the successive transitions (sequential and/or parallel) between three types of audits, namely: traditional audit (type I), ethic and rigor audit (type II), and directed audit (type III). The current framework consider a sequential route, as an audit route that encloses all audit types sequentially executed. Regarding parallel routes, the framework, consider an audit route which registers at least one auditing type repetition (given its poor result). Therefore, the proposed framework allows to track and evaluate audits requirements and results between life-cycle phases (intra-type) and iterations (parallel routes), both within the overall (sequential) route that bound the auditing endeavor. Figure 1 show the proposed framework scope.

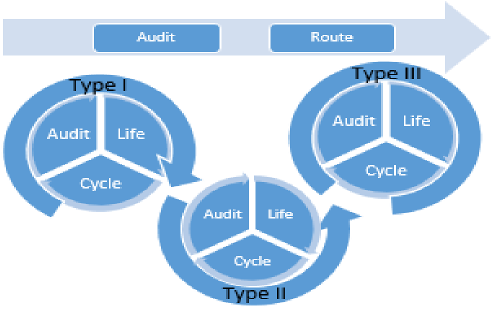


**Fig. 1.** Proposal framework scope

The proposal is not bounded by the cases that contributed, along with existing best practices, to its formulation; the framework, is highly adjustable through a set of parameters in all of its propositions to provide conform with different financial institutions environments. The next section presents the details of the framework.

### 3.1 Audit Life Cycle

One paramount construction is the Analytical Question (QA, the Portuguese acronym). QA is a cast list of questions, defined by the auditing team, targeting what auditors are looking for to inform auditing goals. Auditing life cycle delves with two kind of QA's: preview/planning support analytical question (QAp) and execution Analytical Question (QAe). On one hand QAp consist in a set of QA defined to inform the auditing planning at a macro level, in contrast QAe consist in a set of QA derived from the QAp designed to ensure proper auditing execution. Independently of auditing type, its life cycle consist in the succession of the auditing process phases. Figure 2 shows the envisioned interaction between the phases of audits.

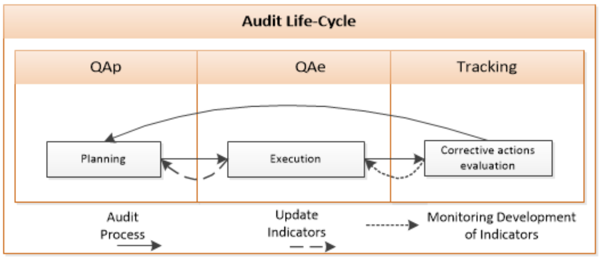


**Fig. 2.** Correlations between audit phases and QA's.

Thus, table 1 show examples of QA´s in the audits life cycle process. The examples are based on one of the demonstrative process, more precisely, term account (DP).

**Table 1.** Analytical question example.

| Phase | Description | Objective |
|---|---|---|
| **Planning (QAp)** | QAp 1) How many DP's were constituted in last 15 days? | Products constituted in a defined times period (is this case DP´s in the last 15 days). |
| | QAp 2) What is the unconformity aggregate of DP's constituted DP´s in foreign currency? | Find out aggregate of unconformity relative to DP´s constituted DP´s in foreign currency. |
| **Execution (QAe)** | QAe 2.1) Relatively the unconfirmed DP´s, how many the applied interest rate are up then in vigor interest rate in the constitution date? | Discovery DP's with unregularly interest rate. |
| | QAe 2.2) For DP´s constituted in foreign currency, the applied exchange rate are in according with exchange rate in effect? | Find out DP's with unregularly applied Exchange rate. |

From the example in table 1 one can acknowledge that, QA´s over the data set that constitutes the Information Base (IB) will drive the focus toward product, services and/or Organic Unit (OU) of the auditing scope. Such is accomplished on a quantitative basis regarding unconformities yielded by the answers to QA's considering the IB on the criteria introduced over this section.

**Definition of audit life cycle indicators.**

*Planning phase: Optimization of action plan.*

The proposed framework provides the mechanism that allow simultaneously: the conversion of the aggregate of unconformities (obtained through QAp) into Key Performance Indicators ($\mathrm{KPI_X}$)[4], and the contrast of these indicators ($\mathrm{KPI_X}$) with the Key Goal Indicators ($\mathrm{KGI_X}$)[5]. Products, services and/or OU which $\mathrm{KPI_X}$ does not comply with $\mathrm{KGI_X}$ become target to the detailed analysis – Execution phase (Figure 3).

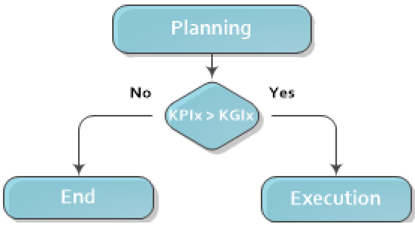


**Fig. 3.** Model to shift products, services and OU.

In the planning phase, for each product and service, the IB is made up of all product and/or service items, contained in the analyzed time period. Therefore, for conversion purposes, is adopted the total items of the product/services as the Base Measure of Product (MBP, being the value of this 100%). For example, it is supposed that have been constituted thirty (30) DP´s in the last 15 days, where the last 15 days is the analyses period. The thirty (30) DP´s represents the base measure (100%) of this product.

Subsequently, the proposed framework provide the auditors team with performance indicators ($KPI_{Mi}$) referent to the aggregate of unconformities of each product or service, as well as performance indicators referent to the aggregate of unconformities of each OU (e.g. a regional balcony) based on the following mathematical expression.

$$KPI_{Mi} = \left(\frac{TIP_i}{MBP}\right) * 100 : MBP \subset [t_n , t_{n+1}[ \tag{1}$$

$$KPI^{*}_{Mj} = \sum_{i=1}^{n}(\%@pp * KPI_{Mi}); \; para \; i = \{1 \cdots Nº \; de \; produtos \; auditar\} \tag{2}$$

While $KPI_{Mi}$ is obtained between the ratio of the total unconformities of the product (TIP, from Portuguese acronym), that is the measure of aggregate of unconformities of a particular product or service, over the MBP; $KPI^{*}_{Mj}$ result from the weighted average of the $KPI_{Mi}$ to a given OU. For such, the $KPI_{Mi}$ is weighted by the %@pp parameter, which is defined by the audit team. Another construct was introduced, is the audit life-cycle key performance indicator ($KPI^{*}_{C}$), that is obtained from the algebraic sum of $KPI^{*}_{Mj}$, as showed in the next expression.

[4] The "KPIx" is used to generically to refer to $KPI_M$ (KPI relative to a product or service) and $KPI^{*}_{M}$ (KPI relative to OU) originally foreseen in the framework, but here aggregated for simplicity and space reasons.

[5] The "KGIx" is used to generically refer to $KGI_M$ (KGI relative to a product or service) and $KGI^{*}_{M}$ (KGI relative to OU) originally foreseen in the framework, but here aggregated for simplicity and space reasons.

$$KPI_C^* = \sum_{j=1}^{n}(KPI_{Mj}^*)\,;\; para\ j = \{1 \cdots Nº\ de\ UO\ auditar\} \tag{3}$$

These overall concept can be explained evoking and expanding the previous example. If in the last 15 days were constituted thirty (30) DP´s, the MBP of this product; and considering that three of thirty DP's are not conformed (e.g. the applied interest rate differs from the rate in effect). In this case, the expression 1, yields 10% of unconformities 10%, in other words, the $KPI_{Mi}$ referent to DP's product is 10%.

*Execution Phase*

The execution phase provide, a fine grain perspective, of specific unconformities of each products, services and/or OU, elicited on the planning phase. Conceptually, the proposed framework attend the execution phase in two distinct and complementary perspectives: the logical perspective and the physical perspective.

*Logical perspective.*

In the logical perspective, for each QAp, are defined subset of QAe. As so, in this perspective, for each aggregate of nonconformities defined in the planning phase, the QAe define criteria to quantify and evaluate specific unconformities inherent to the aggregates of "macro" unconformities. Table 1, exhibit a mapping between cast lists of QA's (QAp 2 and QAe 2.1, QAe 2.2). Therefore, QAe identifies different natures of unconformities in each product/service (INP[6]). Therefore, in detailed analysis (execution phase), for each product and service, the proposed framework identify and provide auditors with overall unconformities of the same nature (TINP).

For performance evaluation, each $TINP_K$ should be converted and delivered to auditors as a performance indicator of the respective $TINP_K$ ($PTINP_K$), that is, indicators that defines, accounting for the weight for the given nature of unconformities, over the aggregate of unconformities of the respective product or service. This indicator is provided by the following expression.

$$PITINP_k = \{[(\%@in * TINP_k) * 100]/TIP_i\} \tag{4}$$

Where %@in reflect the weight (percentage value assigned by audit team) that the aggregate of unconformities of given nature represent over the respective TIP. As an example, if in a universe of 1000 analyzed DP´s in the planning phase there were 500 nonconforming DP's (TIP), and during the execution was verified that the applied interest rate to 200 of them differs of the effect interest rate (TINP, relative to DP's constituted with interest rate differ of interest rate in effect) and the constitution value of 300 of them were less than the minimum amount required (TINP, relative to DP's constituted with amount less than minimum amount required), and also considering that the weight of the TINP relative to DP's constituted with interest rate differ of interest rate in effect is selected as 40% and the weight of the TINP relative to DP's constituted with amount less than minimum amount required is established as 60%, based on expression 4, the PTINP relative to TINP of DP's constituted with interest

[6] In this project the concept of unconformities of the same nature, refer to set of similar unconformities related to a given product or service.

rate differ of interest rate in effect will be 16% and the PITINP relative to TINP of DP's constituted with amount less than minimum amount required will be 36%.

*Physical perspective.*

The physical perspective consist in documentation (that supported transactions) analyses, in order to examine the materialization (or not) of nonconformities detected during the detailed analysis (developed in the logical perspective). The actions, in the field, developed in this perspective may incur in a change/settle over the indicators feed through the logical perspective. Under this conception, the framework provides an equation, named as "regularization expression" over which the previous (logical) performance indicators ($KPI_M$, $KPI_M$* e $KPI_C$*) are updated.

*Phase of corrective actions evaluation.*

In the phase of corrective actions evaluation it is provided the instruments to support the evaluation and tracking of the successive iterations of the audit life-cycle. So, this phase has associated two indicators; one for monitoring support (MC – Continuous Monitoring) and other (IT – Transition Indicator) for iteration/phase transitions threshold definition. The MC indicator records and provide to auditors, the evolution of the performance indicators, during and after successive iteration of the audit life-cycle and the audit route. The proposed framework provide the MC through the following expression (5). Where $KPI_{X+1}$ refer to the $KPI_X$ of the actual audit life-cycle iteration. The ITx indicator is used for performance evaluation. As so this indicator provide variation between $KPI_X$ and $KGI_X$, informing the impact of the performances indicators over the goals indicators. Mathematically ITx is provided by expression 6.

$$MC_x = KPI_{x+1} - KPI_x \qquad (5)$$

$$IT_x = \left[\frac{(KGI_x - KPI_x)}{KGI_x} \times 100\right] \qquad (6)$$

**Performance evaluation and continuity criteria of the audit life-cycle**

Performances evaluation can be of two natures: partial and global. Partial evaluation consist in performance analyze of a specific OU, product or service. The global evaluation is related with the overall audit life cycle performance analyzes as a whole. Independently of its nature, performance evaluation consist in measuring the performances indicators, that is, it is provided by ITx. As so for partial evaluation, the ITx rely on $KPI_M^*$, for OU evaluation, to $KPI_M$, for product/service evaluation, and to $PITINP_k$, to unconformities of specific nature. For global evaluation aids, the ITx fall back on $KPI_C$* (*overall*).

ITx is used as decision criteria of transition between auditing phase's and/or audit (new) cycle (whereas for the same or consequent audit type).

## 3.2 Audit Rout

The audit route concepts, are related with systematization of evaluation and interaction of the three audits types (and their iterations) proposed by the framework. So, the proposed framework recommends that an audit route begins with multiple audits process and during the execution will result in the culmination of the last process with need of repetition.

**Definition of audit route indicators**

The audit route indicators can be partial and/or global. Partial indicator, report the impact that performance indicators, of each cycle of auditing, represent over the route. Therefore, partial indicator of audit route is given by global evaluation of auditing life cycle, that is, the weight of an auditing cycle (PPP) over the audit route is given by the ITx relative to $KPI_C$*. Mathematically PPP is elicited by the following expression.

$$PPP_C = \%@ap * IT_C^* \tag{7}$$

Where %@ap (a parameter - again defined by the audit team) reflects the weight that specific auditing has on the route. It is important to note that for auditing directed to a specific OU, product/service and/or unconformities of given nature the PPP will be given by ITx relative to KPIM*, KPIM and/or PITINPK, respectively. The global indicator (PTP, total route evaluation) provide information for overall evaluation, as such it is obtained from the PPP arithmetic average, as shown in the following expression.

$$PTP = \sum_{c=1}^{n}(PPP_c) / NTA \tag{8}$$

NTA represent the overall cycle of auditing in auditing route. For cases which the audit route is not linear (with parallel courses ramifications), is calculated partial score ($SPI_m$) for each iteration, and then PTP will be the weighted (assigned/established by the stakeholders) average of $SPI_m$.

**Performance evaluation and continuity criteria of the audit route.**

Again, the audit route performance evaluation is of two main fold: partial and global. The evaluation criteria and logic are similar with the previous depicted for the audit life-cycle in the previous section. Nevertheless, it worth to refer that in the audit route are defined indicators that allow auditors team to track and follow up routes of auditing.

## 4 Integrating the framework with multidimensional data modeling

This section define the principles and instruments to couple the concepts proposed on the framework into the multidimensional data modeling paradigm.

### 4.1 Backward compatible management principles

The principles of dimensional data modeling organize the data which allows to answer QA according to the cube metaphor, which is consubstantiated by fact and dimension tables hosted in a DW [18]. Considering that auditors decisions are based on a set of pre-defined QA's, proves to be necessary create a coupling between such QA's and the multidimensional data model that can inform them. Furthermore, over time there might be necessity to include or exclude QA's to/from cast list. So, there is necessity to guarantee the retro compatibility between the interrelations of QA cast list and data model, as referred in [18]. This constitutes pertinent requirement if one is supposed to continuously monitor evolution. Figure 4 show the relation between QA's and the cube of data, and the retro compatibility principle.

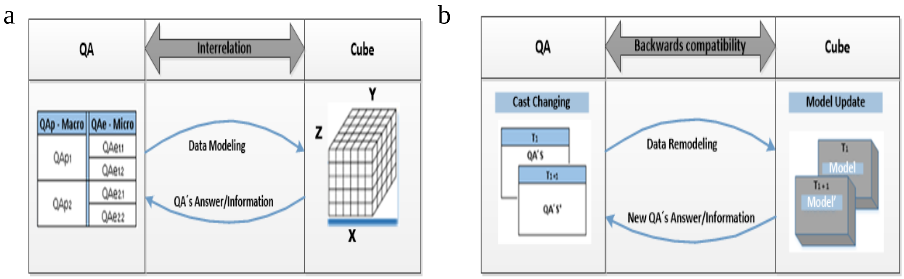


**Fig. 4.** (a) Interrelation between QA´s and cube; (b) Retro compatibility between data model and cast list of QA's.

Figure 4 (b) illustrate that a change occurred in cast list of QA's and then, implicitly, an updated of the cube constitution are required. From this context are proposed, two instruments to manage the retro compatibility consistence and coherence, namely: two-way table and bus matrix.

### 4.2 Backward compatible management instruments

**Two-way table**, defined to this project scope, this instrument consist in mapping, for each audit process, the interception between QA´s and measures (fact table) and attributes (dimension table) of the respective product/service, in order to coherently and consistently update data cubes. **Bus matrix-**in DW system this is an established instrument to conform dimensions, helping the design and the DW architecture project [18]. In this project bus matrix map the correlation between fact tables and dimension tables, that is, includes all fact and dimension table of the model, fostering the conformation.

## 5 Evaluation and discussion

A structured questionnaire was developed, and then administrated to auditing experts/professionals to validate the concepts proposed in the model. The questionnaire was divided in four sections: First section, concerns to evaluate the utility of the proposed key concepts, that is, cycle and audit route; the second section intended to validate concepts related to the criteria proposed to support decision making; the third section aims to elicit the pertinence of the coupling between proposed framework and multidimensional data modeling approach; and the main objective of the fourth section was to obtain comments, suggestion, and constructive critics, which can contribute to improve the model. Altogether the result was satisfactorily positive, on a scale that ranges from 1 to 5 and considering 8 respondents. Comments from the experts stated that the proposed framework is above the average, which suggests that the concepts proposed in the model are useful and applicable.

## 6 Conclusions

Initially, one analyzed the Angolan bank system, it was identified the relevance of adopt tracking and continuous monitoring of audit in bank activities. Grounded on a case study approach directed to three selected bank institutions, operating in Angolan financial market given their representativeness for the purposes of this work. From these cases were identified (good) auditing practices, however low levels of systematization and methodology that improve the efficacy of tracking the audit life-cycle and consequently the audit route, concepts that we develop on the proposed framework.

Therefore, it was defined the criteria which promotes systematic measures to plan, execute and to monitor audit actions. As result it was conceptualized a framework to support the process of tracking audit life-cycle (and audits routes). Though, the proposed framework emerge from illations taken from case study, it is adjustable to bank institutions in and out of case study scope, that is, for adjustability reason, the model provide a set of parameters which are editable by the audit team. Also was considered the requirements for coupling the proposed model with multidimensional data model approach, taking into account the volume and diverse operational data sources for the required analysis. The concepts underlying the framework inform the developed prototype that demonstrate the implementation of a technological tool that has allowed to explore 3 years of data considering the proposed metrics and framework workflows.

For preliminary evaluation a structured questioner was developed and administrated to auditing experts, to validate the model. The result was satisfactorily positive, allowing us to conclude that the proposed framework reveals itself as a useful instrument to support auditing's activities, and its adoption enable auditors with systematic measures that optimize audit action and support decision making.

Although the proposed framework proved to be useful, there are same further work to be done in the future, to improve the model, that is, more precisely: **recommendations management** – with this approach is expected to be developed a methodology to couple and manage in model auditing recommendations; **Data mining approach integration**, the goal here is to enable the proposed framework with some data mining algorithms (for *compliance* activities purpose), to allow recognize patterns under the data, which cannot be identified via QA's; and **proposed framework extension**, regarding its concepts to non-financial institutions, that is, as for example, touristic and/or oil industry.

## References

[1] R. K. Morck and L. Steier, *A History of Corporate Governance around the World : Family Business Groups to Professional Managers*, no. November. Chicago: University of Chicago Press, 2005.

[2] J. Cohen, G. Krishnamoorthy, and A. M. Wright, "Corporate governance and the audit process," *Contemp. Account. Res.*, vol. 19, no. 4, pp. 573–594, 2002.

[3] B. Committee, *Basel Committee on Banking Supervision Consultative Document Principles for enhancing corporate governance*, no. March. 2010.

[4] KPMG International, "Transforming Internal Audit: A Maturity Model from Data Analytics to Continuous Assurance."

[5] KPMG International, "Data Analytics Enabled Internal Audit Survey," 2015. [Online]. Available: http://www.kpmg.com/za/en/issuesandinsights/articlespublications/risk-compliance/pages/2015-data-analytics-enabled-internal-audit-survey.aspx.

[6] A. Tumi, "An Investigative Study into the Perceived Factors Precluding Auditors Form Using CAATS and CA," *Int. Conf. Adv. Res. Bus. Dubai*, vol. 1, no. December, pp. 2–10, 2013.

[7] I. Malaescu and S. G. Sutton, "The Reliance of External Auditors on Internal Audit's Use of Continuous Audit," *J. Inf. Syst.*, vol. 29, no. 1, pp. 95–114, 2015.

[8] ISACA, "Adopting Continuous Auditing/Continuous Monitoring in Internal Audit (Digital)." .

[9] M. A. Vasarhelyi, S. Romero, S. Kuenkaikaew, and J. Littley, "Adopting Continuous Auditing / Continuous Monitoring in Internal Audit," *ISACA J.*, vol. 3, pp. 1–5, 2012.

[10] G. Brennan, "Implementation of a Continuous Auditing System at SIEMENS," 2005.

[11] L. J. Abbott, S. Parker, and G. F. Peters, "Audit Fee Reductions from Internal Audit-Provided Assistance: The Incremental Impact of Internal Audit Characteristics," *Contemp. Account. Res.*, vol. 29, no. 1, pp. 94–118, 2012.

[12] Computer Assisted Audit Group, "A Guide to Computer Assisted Audit Techniques," p. 9.

[13] H. Goyal, "Internal Audit Process," 2012.

[14] I. Pedrosa and C. J. Costa, "Computer Assisted Audit Tools and Techniques in Real World: CAATT's Applications and Approaches in Context," *Int. J. Comput. Inf. Syst. Ind. Manag. Appl.*, vol. 4, pp. 161–168, 2012.

[15] ICAI, *Introduction to CAAT*. ITT FDP, 2015.

[16] P. Dewald, Baya; Hughes, Steve; Turley, *SQL Server analysis services 2012 cube development cookbook*, 1ª Edição. Birmingham: Packt Publishing, 2013.

[17] J. Mei, W. C. Wu, X. D. Gan, and L. B. Liu, "The decision support system of auditing social insurance funds on-line based on data warehouse," *Proc. - 2010 2nd IEEE Int. Conf. Inf. Financ. Eng. ICIFE 2010*, pp. 348–350, 2010.

[18] R. Kimball and M. Ross, *The data warehouse toolkit: the definitive guide to dimensional modeling*, 3ª Edição. Indianapolis: Wiley, 2013.

# Interactive Solution for Multimedia Content Evaluation

Tiago Soares da Costa

INESC TEC, Campus da FEUP, Rua Dr. Roberto Frias, Porto, Portugal,
tiagocosta@inesctec.pt

**Abstract.** In today´s content-centric world, quality evaluation is considered essential to deliver a successful interactive multimedia solution. Set-top boxes, for instance, apart from basic functions such as decoding source signals into content which can be seen by users, also provide the user with a set of tools to evaluate content and decide if it is adequate to his/her preferences. Other platforms even allow the users to provide written feedback and select the most desirable genre for its day-to-day viewing habits. However, most of these solutions rely on traditional control mechanisms, depending on remote controls or keyboard/mouse combinations, leading to a less natural approach to a procedure which content providers would prefer to be more intuitive and less restrictive. In this paper, a novel natural solution for multimedia content evaluation is described, based on tracking hardware, gesture recognition and proven frameworks. By facing a real-life necessity in a given scenario where hand gestures were required for content evaluation, this work delves into natural interactions and their application in common tasks. The end result delivers a new approach to content evaluation, and bridges the gap between this particular field and current gaming-like techniques. These requirements led to an early, careful analysis of interaction techniques and frameworks available and a thorough selection process, which was conducted before any development. The solution translates hand gestures into user interactions within the solution, proceeding according to the user´s expectations. All of this is done within a given time limit, during which the user has to complete the evaluation procedure. The work was tested in a real-life scenario with a limited group of 10 people, the maximum amount achievable under the available circumstances. Each of the users was assigned with a content evaluation task and was required to provide feedback upon the completion of this task. The positive feedback from 6 users from that overall user sample, along with minor suggestions from the remaining users, which rejected the solution, confirms the usability of this work and its applicability for this specific purpose. The tool is capable of being easily used for multimedia content evaluation, allowing users to complete this task without any major hiccups. This was the main milestone set for the development of this solution.

**Keywords:** computer graphics, natural interaction, content evaluation, multimedia content

## 1 Introduction

Quality of multimedia content has become a defining aspect in today´s interactive systems. If in a recent past, the quality of video content used on multimedia systems was not thoroughly scrutinized (leading to the presence of content with questionable quality), its relevance to the overall success of multimedia systems, such as *Netflix* or *Amazon Prime*, was of the utmost importance. Because of this, using ever-improving evaluation techniques applied to multimedia content has gradually become an important factor on state-of-the-art multimedia solutions. Even though the usage of specialized evaluation techniques, applied to selected groups, is still the most preferable solution for the scenarios faced on multimedia systems, using users themselves as a mean to extract feedback is an alternative, interesting and cost-effective solution. For example, applying gaming paradigms to content evaluation and integrating these ideas on an overall solution can lead to an interactive experience where the user itself becomes a part of the solution, is rewarded by this initiative and improves the overall quality of the multimedia content available on the system. On the other hand, seamlessly integrating evaluation mechanisms on new or already existing solutions requires careful planning, with a particular focus on the interaction solutions used for this particular purpose.

Traditional evaluation systems typically rely only on physical controls, such as keyboard and mouse, to gather input information from the users. No additional data is gathered from other sources. Even though valuable information, such as typing speed or keys, can be used to provide emotional state information from the regular users, a complete picture of the content feedback from the user is not achieved with these traditional solutions. On the other hand, *Human Computer Interaction* (*HCI*), a field which was vastly explored during the last decade, enabled interaction between users and computing devices which surrounded him, without any actual physical contact. The use of *HCI* allowed innovative and dynamic approaches to environments, since users were able to rely on pre-defined gestures to complete a set of tasks associated with a given tool. This kind of interaction was popularized on home gaming consoles, where low-cost hardware devices were coupled with interactive gaming applications, achieving relative success. As a direct consequence of the use of *HCI* on modern devices, research on gestural interfaces was intensified. Correlating design approaches with successful interactive solutions led to significant contributions on the development of new interactive graphical user interfaces. However, the lack of established guidelines for gestural control, coupled with the misuse of established conventions, led to poorly designed *Graphical User Interfaces* (*GUI*) which were not adapted to interactive environments [1]. In some cases, they were even adapted from previous solutions, which required accurate interactions and were developed for mouse and keyboard devices.

On the technological front, the launch of *Microsoft Kinect* [2] (or *Project Natal* as was called during development) on November 2010 provided a new approach which was made available to researchers and general public. Combining a camera sensor, infrared projector and a special microchip used for 3D reconstruction, this hardware was capable of delivering depth maps and provide gesture and facial recognition to other applications. Being a low cost solution for interactive systems, enterprises, developers and researchers relied on this solution for multiples purposes, from disease prevention to security probing. In the meantime, more advanced hardware solutions such as *Intel RealSense* [3] have emerged, providing easy to use *APIs* that allow for easy integration on existing projects, while giving developers the means to improve existing tracking algorithms currently available on the hardware.

In this paper, a novel interactive solution for multimedia content evaluation will be thoroughly presented and each of the areas will be explained in detail. Derived from pre-defined guidelines that mandated the development of a new multimedia content evaluation tool, which would rely on innovative forms of interaction to attract users, a new solution was outlined. Based on methodologies traditionally used for content evaluation, this work relies on interactive mechanisms, such as hand gestures, to provide a gaming-like approach and engage users on tasks such as feedback extraction. Independent content categories are available to the users and based on that selection, evaluation will be conducted in real-time. For each video selected for evaluation, the user will approve or reject the content and select a set of keywords which will help to identify that particular video. To provide a gaming approach to this evaluation procedure, the set of videos from each category will be presented within a specified time limit, which will have to be completed by the user. When the user completes the evaluation, results for each video are presented, which can then be saved into the local hard drive. The user can also restart the content evaluation procedure or terminate the content evaluation application. To develop this evaluation solution, *Intel RealSense Software Development Kit (SDK)* and camera hardware were selected for tracking purposes, providing the required tools for a complete, top-to-bottom solution. To deliver an attractive user interface allied with efficient coding and optimal performance, *C#* and *XAML* were chosen as programming languages for this work. Regarding the framework and *Integrated Development Environment* (*IDE*) used throughout this work, *.NET 4.5* and *Visual Studio 2015* were used since both were considered to be the best solutions to develop along with other previously mentioned *Microsoft* technologies.

The rest of the paper is organized as follows. Section 2 will present the current State-of-the-Art on tracking algorithms and systems relying on these technologies for interactive control mechanisms. In Section 3, the proposed approach used for this work will be detailed, along with any necessary requirements and decisions taken during development. In Section 4, the results of this solution will be discussed. On a final note, conclusions and future work are presented throughout Section 5.

## 2 State-of-the-Art

The fast-paced evolution in interactive solutions applied to multiple areas of everyday living leads to numerous research developments that have contributed with valuable knowledge. In [4], a natural interaction solution was developed along with a support graphical interface. Relying on human gestures for *User Interface (UI)* interaction, the solution was able to replicate current day-to-day user experience, based on mouse and keyboards, into a more intuitive experience. Since the main goal of the authors was to replace the mouse as preferential mean of interaction, this was accomplished with vast success through the use of low-cost hardware such as *Microsoft Kinect*. Input frames are pre-processed: background information is subtracted, images are segmented and human contours are tracked. Based on this information, the manual gesture is identified and the system proceeds with the according action, mapped previously by the authors. The solution presented high success rates regarding the gesture identification, even though some improvements are still required on the detection algorithm to improve its performance.

On [5], a gesture-based computer mouse solution was presented and tested by the authors. Using a *Kinect* sensor, an interactive system to replicate the mouse control was developed. Being a comfortable and reliable solution, this was an easy-to-use interaction mechanism, even by unexperienced users. Data such as precision and speed from hand movements were used as parameters for this solution, leading to an efficient tracking and gesture recognition algorithm which delivered convincing performance results, especially when compared to a physical mouse interaction. The authors translated hand movements into cursor movements, which resulted into a “joystick-like” user interaction. While the majority of data available on a given frame was discarded, hand and finger positioning was used on the detection algorithm to identify the correct gesture. Detailing this procedure, the algorithm uses the top of the fingers, along with the indicator and the thumb for this particular procedure. Results have shown that, even though this approach led to results that were comparable to traditional mouse interactions, initial discomfort was the main issue that affected users. To overcome this problem, an initial tutorial or experimental period should have been given to users, so that they could have been familiarized with the system before any testing was conducted.

A new approach for interactive medical solutions was presented on [6], with considerable success. Using *Microsoft Kinect* as a mean to gathering visual information, a *GUI* system was specifically developed for intensive care in hospitals. A simple graphical user interface, prepared for interaction through hand tracking, was used on this particular solution. No additional sensors, gloves or specific colors are required on this system. The authors rely on skeleton information provided by *Kinect* to extract the hand coordinates from the hands and detect any click on a given point in space. If required, the solution can interact with the medical staff or auxiliary systems through auxiliary communication mechanisms such as email, telephone, voice or *SMS*. Field tests were conducted and the solution was approved by regular users and medical staff, both in response timings achieved on day-to-day situations and easiness of usage of the provided graphical user interface. However, the basic *GUI* that was used on this solution, a less than ideal selection of interaction elements such as radio buttons, a rather simplistic approach presented throughout this work and inconclusive test results have shown that further work is necessary in order to improve the quality of this solution.

Gaming interaction was the main topic of research on [7], where an interactive car game based on gesture recognition was presented. A system based on *Microsoft Kinect* was developed and integrated into a car game, using the skeleton information as a mean of control of the vehicle. A set of 4 gestures was identified and used for interaction between users and game and the underlying delay associated with this type of scenario was taken into account during the development of the solution. The system tracks left and right hands, along with the shoulder position and translate that specific information into the corresponding gesture, selected from a predefined set of gestures mapped beforehand. That gesture is then interpreted by the videogame as a specific action, which is reproduced in real-time with minimal delay. Exhaustive tests conducted on this solution validated the application of this interactive approach on a virtual, action-based scenario such as interactive gaming. However, since pre-stored posture information is required beforehand, the use of this solution on different contexts requires additional pre-processing steps. This fact hinders its usefulness, despite the performance achieved by the algorithm presented on this work. A dynamic solution, capable of adapting to multiple contexts, would have been desirable.

On [8], a new algorithm to improve the performance of hand contour tracking is extensively presented. Real-time depth information is extracted from a *Microsoft Kinect* input camera. The new solution surpasses the performance achieved by solutions currently available on the market. This new approach gathers depth information and identifies the nearest pixel from a given input frame. The contours and palm of the hand, as well as the position of the fingers, is identified in real-time by the solution. After pre-processing this information, the correct gesture is identified and the outcome of this task is sent directly into the application. The success rate achieved was approximately 94.6%, a result that approves the work done on this solution regarding dynamic and static gesture recognition. Despite the successful results achieved with most of gestures, some specific gestures (like I, N or T) have presented inferior results and require additional research to improve its recognition rates.

Natural interaction with television (*TV*) content was the main topic of research on [9], which presents a complementary solution for interactive television. Relying on *Microsoft Kinect* software and hardware for dynamic gestures and handwriting recognition, the authors achieve a solution capable of selecting *TV* channels and interact with a *TV* set through a pre-defined set of actions from the users. For this purpose, the solution relies on the skeleton data provided by *Kinect*, from which left hand and right hand information is extracted. While the left hand is used for interactive *TV* operations, the right hand is used for handwriting procedures within the solution. Based on the hand information from the user and the respective gesture executed, the system identifies the appropriate action and proceeds with the corresponding result. After testing, the results achieved by handwriting and gesture recognition were considered positive. However, the effect of the error threshold defined within the solution on the recognition rate was significant. Due to this, the definition of this specific value required particular attention, in order to avoid any negative impact on the performance achieved throughout execution. Alternative recognition mechanisms should be analyzed to solve this performance problem.

When *Microsoft Kinect* was presented to the world, a cost-effective solution for tracking and recognition, without any particular accessory requirement, was made available to the public. However, other tracking solutions which rely on accessories for tracking are still currently available on the market. A solution of this type is detailed in [10], with encouraging results based on real life scenarios. An alternative approach to current tracking solutions, based on wearable devices was proposed and information such as gestures and rotation were used as control mechanisms. The system identifies the curvature, direction and gesture of the user, based on the data provided by the sensors. From that information, the system translates data into actions, just like any user would expect with regular mouse and keyboard interactions. Despite the positive tracking results, drawbacks such as requirement for an additional device and lack of optimization of the solution lead to a mixed reaction towards this solution. Touchless interaction without additional devices is still the preferred approach for these type of solutions.

User emotions were the focus of attention of the work presented in [11]. An adaptive user interface was presented, allowing interactions between users and the provided system. Keyboard strokes, application events, emotional state and color selection were used as means of identification of the appropriate state of the user within the *Ortony, Clore and Collins* (*OCC*) model. Only 3 types of emotions were considered for this particular work: satisfaction, agony and anger. Based on the writing speed, keys used on the keyboard and events of the system, the user interface is dynamically changed according to this input data. Experiments with this solution confirm its use, without the need for any visual references as input data. However, the authors should have conducted extensive testing, in order to analyze user´s emotions when dealing with this dynamic solution.

Facial recognition information was used for emotional state tracking on [12]. Gathering depth data extracted from *Microsoft Kinect*, the authors were able to analyze facial data and recognize current expressions from users. From the face position and references points provided to the solution, the system was able to select an emotional state from a complete set of 9 emotions. To achieve this, input frames were pre-processed (due to frame-to-frame inaccuracies) and support data such as occupied area, depth and face orientation was collected. For facial references, the system tracked 3 specific regions: chin, cheeks and eyebrows. The testing results showed the applicability of this solution on real life scenarios, but complementary data was deemed necessary for additional confirmations. Problems observed with the detection of certain expressions have to be taken into account and a thorough analysis of mandatory facial information is required. Only then will the authors be able to deliver a complete solution, capable of accurately identify emotions regardless of the subject used for testing.

The State-of-the-Art articles detailed throughout this section present new approaches to problems faced in multiple interactive solutions. However, none of these articles actually dealt with the applicability of interaction mechanisms to scenarios such as multimedia content evaluation, a task typically associated with physical devices such as keyboard and mouse. This opportunity became the main milestone for our work, as can be seen on the following sections.

## 3 Proposed Approach

For this work, a *Multimedia Content Evaluation Application*, capable of allying gaming techniques with audiovisual content evaluation, was proposed. The motivation for the development of this solution was found when a previous analysis of currently available content evaluation tools was conducted. Although several commercial or open source tools were available for use, their interactions relied solely on mouse and keyboard events. This presented us with a problem and a research question: how to adapt an already existing task to new interactive environments, relying on control mechanisms commonly used by users. Besides interactions requirements typically laid out when a new solution is developed, other complementary aspects (such as performance or user experience) had to be analyzed. Since the content, which would be evaluated with these tools, was going to be used on an already existing interactive solution, a goal was set for this work: harness this opportunity and expand these interactive features to such common tasks. An initial approach was outlined for this work: research current trends and select the most efficient interaction solution, design the *UI* and use interactive mechanisms required to successfully complete this task.

Starting with the *UI* design, current evaluation tools aren´t adapted to touchless environments, such as the one that would be used during this work. They are usually based on navigation bars and menus, a design approach which isn´t practical when applied to environments without physical interaction devices. This led to the application of design guidelines, commonly used on other *UI* solutions, for an effortless interactive experience. The use of elements, such as buttons, benefited from contributions which defined the most adequate settings for an accurate experience. Based on that knowledge, pre-defined interactions between users and *UI* were defined within a given set of screens, each with a specific task which had to be completed beforehand to proceed into the following screen. Regarding the technical development and integration of the solution, the work was focused on a modular application, capable of integrating multiple sub-modules for each corresponding area while delivering a complete and functional package.

Based on all of the requirements set previously, an overall solution was specified and split into 3 specific modules, as represented in Fig. 1*: RealSense Tracking*, *Content Evaluation Mechanisms* and *User Interface*. All of the development work was conducted in *C#*, using *Microsoft Visual Studio 2015*. Regarding the *UI*, a single, centralized application screen, with all *UI* components contained within a single *XAML*[1] file, was devised. Relevant *UI* information would be shown or hidden according to user interactions. Resource usage would be optimized since the application would load all resources only once to the computer´s memory. A total of 4 application *UI* areas were identified for the *Multimedia Content Evaluation Application*:

- *Category Selection*
- *Multimedia Player*
- *Content Evaluation / Keyword Selection*
- *Results presentation*

Aside from the user interface, a new mean of interaction between users and *UI* components was selected. Instead of the traditional keyboard and mouse controls, the approach on this work relied on gesture tracking to detect hand positioning and determine whether or not the user was issuing a command to the user interface and if that

[1] *XML*-based markup language developed by *Microsoft*, for user interface design within their own commercial products or solutions. It is available in *.NET Framework 3.5*, as an integral part of the *Windows Presentation Foundation*

was the case, which type of command. To do this, *Intel RealSense* hardware and *SDK* was selected for this specific purpose.

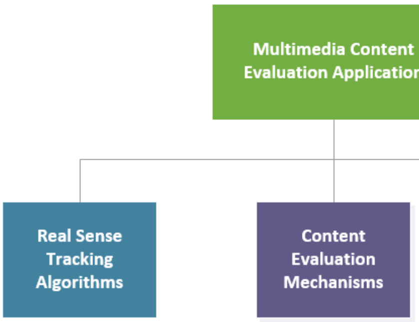

Multimedia Content Evaluation Application
Real Sense Tracking Algorithms
Content Evaluation Mechanisms
User Interface

**Fig. 1.** Software modules which form the content evaluation tool

The new technology, made available by *Intel* to *Original Equipment Manufacturers* (*OEM*) and developers, allows natural interaction between a user and the machine which is being controlled. A small package composed of a regular *1080p* camera, an infrared sensor and a laser projector is available for purchase and depending on the purpose of the application, a short or large range camera can be selected. Besides the hardware, *Intel* also provides an easy-to-use *SDK*, capable of being used on multiple programming languages such as *C++*, *Java* or *C#*. Tracking algorithms can be used or even improved by developers, depending on the goal they want to achieve with their own applications. *Intel RealSense* has been applied to multiple scenarios such as augmented reality, immersive experiences and gaming demonstrations. For this work, tracking algorithms which relied on *Intel RealSense* technology were executed on a parallel thread, allowing the application to capture interaction information in real-time without performance degradation on the user interface thread. This was the case on a preliminary test conducted before any development work had been executed. To provide the required touchless features for the application, the *Touchless Controller* component from *Intel RealSense* SDK was used and depending on the gestures of a given user, multiple corresponding actions were coded:

- If *Tapping* is detected, translate into an *Selection* key press event
- If *Swipe* is detected, translate into an *Start/Stop* key press event
- If *Thumbs Up* is detected, translate into an *Deselect* key press event

The key mappings were achieved by using *Gesture Mapping*, which allows the application to automatically translate input features into whatever action is required by the developer. To prevent any unexpected interaction apart from these coded gestures, keyboard and mouse interaction was disabled by default on the *Multimedia Content Evaluation Application*.

Regarding the *Category Selection* screen, a central grid composed of 3 buttons was outlined and 3 content categories were selected for this proof of concept: *Automotive*, *Adventure* and *Health*. Since the purpose of this type of application would be to evaluate content that would be used on multimedia applications, these themes were selected due to their large variety of content readily available. Their significance on today´s world was also relevant to this decision, since these are some of the most searched topics on online platforms such as *Youtube* or *Vimeo*.

For category navigation on this first screen, swipe gesture was introduced. Raising any of the hands from the user automatically starts the hand tracking procedure, which will then detect if the user is swiping left or right. By doing these gestures, the users are able to navigate through the available categories and select the one that interests them. Selection is done by tapping (moving the hand forward as if pressing an object) the category button and to encourage the navigation between categories, only the central category is available for selection at any given moment. After the selection of the category, the *Multimedia Player* screen is presented to the users.

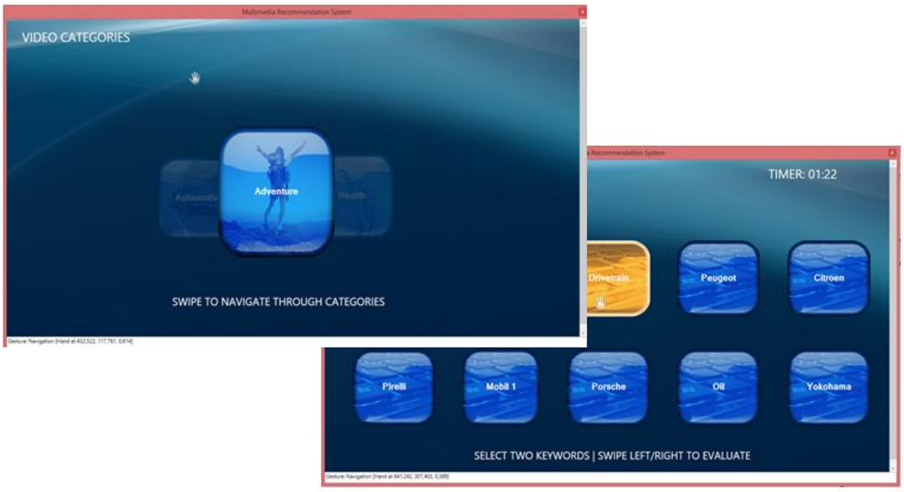

Multimedia Recommendation System
VIDEO CATEGORIES
Automotiv
Adventure
Health
SWIPE TO NAVIGATE THROUGH CATEGORIES
TIMER: 01:22
Drivetrain
Peugeot
Citroen
Pirelli
Mobil 1
Porsche
Oil
Yokohama
SELECT TWO KEYWORDS | SWIPE LEFT/RIGHT TO EVALUATE

**Fig. 2.** Example of the content evaluation tool interaction screens

For content playback, the *MediaElement* was used on the *GUI* design. This object allows the application to modify and reproduce multimedia content in real-time, from a wide variety of codecs supported such as uncompressed *RGB*, *Windows Media Video*, *H264* and *MPEG-4 Part 2.* From a pre-defined set of multimedia content, based upon the category selected previously, the application randomly selects one video for playback. A timer is then presented on the top of the screen, starting a countdown procedure which encourages the user to complete the evaluation within the available timeframe. The title of the video is also presented on this area, so that users can identify the content that is being presented to them. To proceed to the content validation screen, a thumbs up gesture is required since this action is mapped into the *Deselect* key press event, which will hide actual user interface components, such as the *MediaElement.*

For content approval or rejection, the user is required to swipe right or left respectively, gestures which are mapped to *Start* or *Stop* key presses. Timed pop-ups are shown after each content evaluation from the users. These swiping gestures are done within the *Content Evaluation* screen, where a grid of 10 buttons with randomly selected keywords is presented. For each category available on this application, a set of 20 keywords is provided for this random selection. The buttons are available for tapping by the user, which is translated into the corresponding select action. For each video, a maximum of 2 keywords must be selected. These selections, along with the content validation through swiping, are the required actions to proceed to the next video playback. The video playback and content evaluation procedures are sequentially repeated until all videos from the provided set were played to the users or if the time limit presented to the users has ended.

The final screen presented by the application is related to the results of the content evaluation. Composed by a grid of 3 buttons and a single *Text Block*, this screen is significantly relevant for both users and content providers. Since these results are fundamental for both parties, all required data is directly presented on the screen: title of the video, evaluation of the video and keywords selected are printed into the *Text Block* element. These results can also be saved into a text file on the local hard drive, by tapping the provided *Save Results* button. Besides this action, the user can restart the content evaluation through the *Restart* button or, on the other hand, terminate the application by reaching the *Exit* button. An example of multiple screenshots from the *Multimedia Content Evaluation Application* can be seen on Fig. 2.

## 4 Results

To validate the design solutions and implementation of the application, a test in a real life scenario of the completed solution was planned and conducted. To achieve this, the *Centre for Telecommunications and Multimedia* from *INESC TEC* was chosen as the most adequate location for this testing phase, taking into account hardware transportation requirements and user´s availability. Within this location, a group of 10 people from different working areas (such as *Music*, *Video* and *Networks*) was selected, so that the results were not biased. Each person was responsible for the category selection, content evaluation and general feedback from the application experience. Regarding the group of users selected for these experiments, all of the users had academic background, ranging from *MSc* to *PhD* and were up-to-date with recent technological advances. Ages were comprised between 26 and 35 and a random number of male and female individuals were chosen. This guaranteed a homogenous group, which would contribute with accurate results, since there were less variables to interfere with.

The user feedback from the *Multimedia Content Evaluation Application* was mostly positive, with 6 users liking the solution in its current form and demonstrating interest in using it in other scenarios, such as music content evaluation. One user did not like the application in its current form, but would use it if some minor modifications were done to the interaction mechanisms, such as replacing some gestures with more user friendly actions. Finally, 3 users firmly disliked the application and would not use it in other contexts, even with modifications.

Besides the previously mentioned testing tasks which were assigned to them, suggestions on possible applications for this solution were also asked to the users. From all the scenarios presented, the main interests resided on the integration in mass market solutions or within a specific user experience. For example, game reviews on home consoles such as *Playstation 4* or *Xbox One*, or evaluation of *TV* shows presented on set-top boxes were some of the examples suggested by users. Cinemas were another possible scenario for this application: user reviews with no filtering or bias applied could be gathered right at the end of movie presentations, allowing producers to collect valuable information directly from the most important source for their business.

## 5 Conclusions and Future Work

By applying existing interaction mechanisms to an otherwise traditional task, this work was able to attract new users and involve them with multimedia content evaluation. The proposed approach focused on providing new interactive features to an existing scenario, while seamlessly providing feedback for multimedia content that would be used for specific purposes. All of this was achieved without the use of any particularly complex or expensive hardware and relying on currently available *Software Development Kits*, adapted to the required goals within this specific project.

Integrating *Intel RealSense* technology on this work opened up a wide range of opportunities for tracking purposes, such as the hand tracking used throughout this application. With the use of this technology, we were able to introduce gesture recognition and deploy a well though interaction scheme that was both natural and easily remembered by users. Gestures such as thumbs up, swiping or tapping are common to many other current day applications such as gaming or interactive *TV* sets. Using them on this work allowed us to take a common knowledge base, already established between users, and apply it to another context with vastly different goals.

Content evaluation is an often overlooked task, but its importance within commercial multimedia solutions grows every day. While new products showcase their technical advances and improved user features, the general public focuses on available content and whether or not they should subscribe to a given service. Based on this demand, content evaluation is an important tool to understand user preferences and determine if the content choices which have been made meet the user´s expectations.

By integrating the technological aspect with content evaluation purposes, the *Multimedia Content Evaluation Application* was able to combine both worlds and deliver a useful and interesting tool, which can be deployed with relatively small effort on wide variety of scenarios. The tests completed at *INESC TEC* confirmed the actual concept of this solution and presented new challenges for the road ahead. Modifications and suggestions will be taken into consideration and thoughtfully implemented in the future, while retaining the overall concept which was approved by the vast majority of users. The next step in the development of the *Multimedia Content Evaluation Application* will focus on its integration in an overall solution, designed for adaptable *3D multi-view* content streaming, where it will be used to select the most appropriate content. This will allow us to provide the most immersive experience possible to the users.

**Acknowledgments.** This work has benefited from the remarkable support and technical conditions provided by the *Centre for Telecommunications and Multimedia*, located at *INESC TEC*, throughout this research period. On a personal note, I would like to express my gratitude to my senior supervisor, Prof. Maria Teresa Andrade, for the valuable help regarding the proposal and development of this research work.

## References

1. Norman, D.A., Nielsen, J. (2010). Gestural Interfaces: A Step Backwards in Usability. Interactions, vol 1 (ACM, September 2010), p.46-49.
2. Microsoft Kinect (2016) Kinect for Windows. http://developer.microsoft.com/pt-pt/windows/kinect. Accessed 16 Dec 2016
3. Intel RealSense (2016) Technology Overview. http://www.intel.com/content/www/us/en/architecture-and-technology/realsense-overview.html. Accessed 16 Dec 2016
4. Siddharth, S.R., Anupam, A. (2012). Design of gesture recognition system for dynamic user interface. 2012 IEEE International Conference on Technology Enhanced Education (ICTEE).
5. Szeghalmy, S., et al. (2014). Gesture-based computer mouse using Kinect sensor. 2014 5th IEEE Conference on Cognitive Infocommunications (CogInfoCom).
6. Raheja, J. L., et al. (2014). GUI system for elders/patients in intensive care. 2014 IEEE International Technology Management Conference.
7. Pradeep Kumar, B. P., et al. (2015). Gesture recognition based interactive car game. 2015 International Conference on Electrical, Electronics, Signals, Communication and Optimization (EESCO).
8. Plouffe, G., et al. (2015). Natural human-computer interaction using static and dynamic hand gestures. 2015 IEEE International Symposium on Haptic, Audio and Visual Environments and Games (HAVE).
9. Sheu, J. S., et al. (2014). Design an Interactive User Interface with Integration of Dynamic Gesture and Handwritten Numeral Recognition. 2014 International Symposium on Computer, Consumer and Control.
10. Fukui, R., et al. (2015). Hand gesture interface for content browse using wearable wrist contour measuring device. 2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS).
11. Dalvand, K. and M. Kazemifard (2012). An Adaptive User-Interface Based on User's Emotion. 2012 2nd International eConference on Computer and Knowledge Engineering (ICCKE).
12. Szwoch, M. and P. Pieniążek (2015). Facial emotion recognition using depth data. 2015 8th International Conference on Human System Interaction (HSI).

# SESSION 4

## MACHINE LEARNING & DATA MINING

**Coreference Resolution in Portuguese: The Impact of Training Set Generation Approaches**
*Gil Rocha and Henrique Lopes Cardoso*

**Comparison of Classification and Regression Approaches for Predicting Age of Onset**
*Maria Pedroto*

**Combining Forecasters using Arbitration for Water Consumption Forecasting**
*Vitor Cerqueira and Luís Torgo*

# Coreference Resolution in Portuguese:
## The Impact of Training Set Generation Approaches

Gil Rocha[1] and Henrique Lopes Cardoso[1,2]

[1] DEI/FEUP, Faculdade de Engenharia, Universidade do Porto, Portugal
[2] LIACC, Laboratório de Inteligência Artificial e Ciência de Computadores, Universidade do Porto, Portugal
{gil.rocha, hlc}@fe.up.pt

**Abstract.** The aim of coreference resolution is to automatically determine all linguistic expressions presented in a piece of text that refer to the same entity. In this article, we present a preliminary approach to coreference resolution from text written in Portuguese based on the mention-pair model. We also report on the impact that different methods to create the training set from a manually annotated corpus have on the performance of the system. Given the dataset used in our experiments and from the critical analysis of the obtained results, we conclude that semantic-based features will be required to improve the results presented by the system, different methods to create the training set have an high impact in the quality of the obtained results and, that the results presented in this article can be improved if an annotated corpus containing a significantly higher number of annotations is available.

**Keywords:** Natural language processing, machine learning, coreference resolution

## 1 Introduction

Coreference resolution is a natural language processing (*NLP*) task that aims to automatically determine all linguistic expressions presented in a piece of text that refer to the same entity. A natural language expression used to perform a reference is called a *referring expression* (or *mention*), while the entity that is being referred is called the *referent* (e.g. person, company, real-world event).

Automatic recognition of coreferring expressions for entities and events has been well recognized in the natural language processing community as an important step in order to understand natural language, with high impact in several sub-tasks that are currently being addressed by the community. Coreferences are typically associated with natural language ambiguity present in text documents, and are commonly reported as one of the major sources of errors in several NLP sub-tasks. The use of coreferences is prevalent in natural language texts, and occurs in varied ways; oftentimes different coreferences to a same entity are found in a single text. It is well known that writers tend to avoid repetition of words, making use of different referring expressions, and tend to omit implicit knowledge (assuming that it is commonsense), with the aim of obtaining a more fluent reading experience and capture the reader attention. Instead, they often appeal to the knowledge about the world and inferring capabilities they assume the target reading audience has, to convey information about the world. These assumptions turn out to pose very difficult challenges to computational systems aiming to automatically process and reason about information expressed in natural language texts. As an example of the task that we aim to address in this work, consider the following passage (extracted from Jurafsky and Martin [9]):

> "Victoria Chen, Chief Financial Officer of Megabucks Banking Corp since 2004, saw her pay jump 20%, to $1.3 million, as the 37-year-old also became the Denver-based financial-services company's president. It has been ten years since she came to Megabucks from rival Lotsabucks."

In this passage, each underlined pieces of text are used by the author to denote one entity in particular (*i.e.* the person named "Victoria Chen"). We refer the use of these linguistic expressions, like "her" or "Victoria Chen" that are being used in the text to denote an entity as *referring expressions*. The set of referential expressions that can be found in natural language is quite rich indeed, where referential phenomena can be used to express new or given information. Different types of coreferences were identified and are extensively studied in linguistics. The classification between different types of coreferences can be made taking into consideration at least the following dimensions: (a) Positional view: taking into account the relative position in which the referring expression occurs within the text and the corresponding referent; and (b) Syntactic view: taking into account the syntactic function of the *referring expression*.

In terms of to the positional view we can classify a coreference as:

- *Endophora*: the referring expression refers to a referent that is explicitly present in the text:
  – *Anaphora*: the referent was previously introduced in the discourse (a typical example is the use of pronouns, such as "she" or "her", as referring expressions);
  – *Cataphora*: the referent is presented after the occurrence of the corresponding referring expression (e.g. the pronoun "he" in the following excerpt: "When he arrived home, John went to sleep.")
- *Exophora*: the referring expression refers to a referent that is not explicitly presented in the text (e.g. "That chair is broken.")

Referring expressions can also be classified in different types taking into account their syntactic role in a sentence. Based on the classification provided in [9], we can classify *referring expressions* into the following types: Proper nouns (*e.g. John*), Pronouns (*e.g. he*), Definite noun phrase (*e.g. the president*), Indefinite noun phrase (*e.g. a man I know*) and, Demonstratives (*e.g. this man* or *that man*).

A *coreference chain* (also denoted as *entity*) is a group (or cluster) of referring expressions that refer to the same referent (*i.e.* linguistic expressions that corefer). Thus, a coreference chain is associated to an entity and is formed by all mentions in a discourse that refer to this entity [13].

For the passage previously presented, the complete set of possible coreference chains is the following:

1. {*"Victoria Chen", "Chief Financial Officer of Megabucks Banking Corp", "her", "the 37-year-old", "the Denver-based financial-services company's president", "She"*};
2. {*"Megabucks Banking Corp", "the Denver-based financial-services company", "Megabucks"*};
3. {*"her pay"*};
4. {*"Lotsabucks"*}.

Given a text document as input, the goal of a coreference resolution system is to determine all the coreference chains presented in the natural language text. Indeed, the task of coreference resolution corresponds to a very important pre-processing step for several *NLP* subtasks that require natural language understanding, such as document summarization, question answering, information extraction, and argumentation mining. By automatically detecting coreferences in a given text document and being able to aggregate different mentions in the same coreference chain, we are able to disambiguate some of the meaning of the text and, therefore, contribute to a better understanding of the text.

This paper is structured as follows: Section 2 presents some of the related work on coreference resolution, focusing on text written in the Portuguese language. Section 3 introduces the corpus that we have used in our experiments to validate the approach presented in this work. Then, in Section 4 we describe the methods that were used to address the task of coreference resolution from text. Section 5 presents the results obtained by the system described in this work. Finally, Section 6 concludes and points to directions of future work.

## 2 Related Work

General and complete coreference resolution systems receive as input text written in natural language and return as output the set of coreference chains present in the text. Typically, this process is divided in three main components (as defined in [13] and depicted in Figure 1): *mention detection*, *characterization of mentions*, and *resolution*.

The first component of the coreference resolution process is the detection of mentions from text written in natural language, also known as *Named Entity Recognition* (NER) in the *NLP* community. The aim of this component is to find the exact boundaries of each mention (i.e. referring expressions) present in the text document.

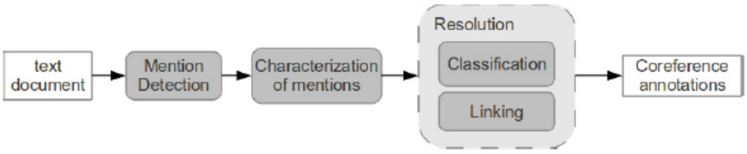


**Fig. 1.** General Coreference Resolution System Architecture (Sapena *et al.* [13])

In the characterization of mentions step, identified referring expressions are characterized with all the available knowledge about them and their possible relations with other mentions. This component is very dependent on the approach being followed and may involve the extraction of lexical, syntactic, structural and semantic knowledge of each referring expression and/or set of referring expressions.

In the literature several paradigms to address the task of coreference resolution have been explored: rule-based, heuristic-based (e.g. based on the centering theory of the discourse [8, 15]), based on constraints satisfaction (e.g. [12]) and machine learning. Since in this article we aim to address the task of coreference resolution using a machine learning paradigm, the remaining of this section will focus on related work following the machine learning paradigm.

The models found in the state-of-the-art for machine learning approaches to coreference resolution can be divided in: mention-pair models, mention-ranking models and entity-mention models.

Classifiers based on the *mention-pair model* are trained to learn whether two referring expressions (or mentions) corefer or not. Therefore, in this model the coreference classification task is typically formulated as a binary classification problem: for each mention pair the classifier returns the target class (CO or NC). Taking into consideration the approach followed in the mention-pair model, the feature vector extracted for each pair of referring expressions should represent the relation between the two referring expressions (characterization of mentions component).

*Mention-raking models* are trained to learn if two referring expressions corefer or not (similarly to the mention-pair model), but for each candidate pair of referring expressions the classifier takes into account the entire candidate set modeled in the form of conditional probability [2].

Finally, *Entity-mention models* formulate the problem of coreference resolution as a clustering problem, in which the classifier aims to group together all the mentions referring to a particular entity in a single cluster.

In [6], Garcia *et al.* proposes a coreference resolution approach that is tailored for entities of the type person, entitled *Link-People*. The corpus used in their experiments was built by the authors [7]. Their model combines a multi-pass architecture and a set of constraints and rules. Some sets of rules are specific to deal with pronouns, anaphora and cataphora for person entities. They mention that their results could be improved if the model was able to detect semantic relations like synonymy, hyponymy and hyperonymy.

Coreixas [1] proposes a coreference resolution system for Brazilian-Portuguese based on a machine learning approach and focusing on semantic categories of named entities, such as *Person*, *Organization*, *Event* and *Other*. Resources used include the *HAREM* corpus [5] for semantic categories, the *PALAVRAS* parser and the *Summ-it* corpus [3] for coreference resolution. Coreixas showed that the approach using semantic categories of entities provided improvements in terms of performance to determine whether a pair is coreferent or not comparing to a baseline approach without semantic information. Also, Coreixas emphasizes the importance of world knowledge for this line of work, since some semantic categories such as *Event* and *Organization* did not show satisfactory results.

Finally, Fonseca *et al.* [4] proposes a supervised machine learning system to address coreference resolution for text written in Portuguese-Brazilian language following the mention-pair model.

## 3 Data

In order to successfully apply supervised machine learning algorithms, a set of labeled data (corpus) is crucial. Therefore, the availability of a collection of annotated documents (corpus) with coreference chains occurring in natural language textual resources written in Portuguese is an important requirement in order to address the task of coreference resolution for Portuguese.

To the best of our knowledge, current available corpora annotated with referring expressions and coreference chains from text written in the Portuguese language are the following: (a) Garcia *et al.* corpus [7]: contains 91 journalistic documents with approximately 34000 tokens and 6 encyclopedia documents with approximately 17000 tokens. The coreferential annotations were manually made by two linguists following the SemEval 2010 Task format. These annotations are restricted to persons and pronouns. Specifically for the Portuguese language, this corpus contains texts from Portugal, Brazil, Mozambique and Angola. The authors claim that having annotations from text documents extracted from different Portuguese language dialects is beneficial regarding the orthographic agreement changes that are being done to the Portuguese language; (b) *HAREM* corpus [5]: contains manually annotated named entities distributed in ten semantic categories. Relations between these entities have also been manually annotated in four types: identity, inclusion, placement and other. However, since *HAREM* refers only to named entities, other noun phrases referring to the entities (e.g. "the president") are not annotated in this corpus. It contains approximately 887 coreference chains annotated from 129 documents written in the Portuguese language. However, a detailed analysis of the annotations related to the relations between referring entities contained in the *HAREM* corpus showed that some of these relations were annotated considering semantic relations between the named entities instead of coreference relations as described in previous sections of this article and; (c) Summ-it++ corpus [3]: consists of fifty journalistic articles from the Science section of the Brazilian newspaper entitled "Folha de São Paulo". The corpus contains a total of 560 coreference chains (manually annotated) with an average of 3 referring expressions in each coreference chain. The largest chain contains 16 referring expressions. These annotations were obtained from newspaper articles written in the Portuguese (Brazilian) language, with approximately 20000 tokens in total. The annotation of coreferences follows the *SemEval* style: the coreference relations are shown in the last column in a numerical bracketed format, in which every referent has an unique ID number and every referring expression is marked with the ID of the referent it refers to.

We have used the *Summ-it++* corpus to validate the approach described in Section 4. Even though this corpus corresponds to the linguistic resource containing the lowest number of text documents annotated for this task, it appears to be the corpus (for the Portuguese language) that has the most reliable annotations, according to a manual analysis.

The *Summm-it*++ corpus follows the typical annotation format used in corpora provided in previous editions of the *NLP* challenges *SemEval* and *CoNLL* related to the task of coreference resolution, in which each data instance corresponds to a complete text document (e.g. a complete news article) that was annotated and presented in the corpus in a tabular format, where each line corresponds to the tokens extracted from the text and each column corresponds to the annotations that were made at the token level. Specifically in the case of the *Summ-it*++ corpus, the annotation file is structured in the following format: (a) each line contains one of the tokens extracted from the original text, and (b) each column contains information related to each token, namely: the word lemma, part-of-speech tagging, gender, number, head word indicator, semantic category (comprising the semantic classes of *Abstraction*, *Event*, *Organization*, *Other*, *Person*, *Place*, *Thing*, *Time*, *Value*, and *Work*), relation descriptor between a pair of named entities, and coreference indicator. In addition, since the exact boundaries of each referring expression are already provided in the annotation file, this corpus provides the necessary resources to focus on the resolution component and assume that the expected output of the mention detection component is given as input.

## 4 Method

To address the task of coreference resolution from natural language text written in the Portuguese language we aim to employ supervised machine learning techniques. In previous related work, these techniques have shown promising results. From the general coreference resolution system introduced in Section 2, we assume that the output of the *mention detection* component is given as input and focus our attention on techniques and methods for the *characterization of mentions* and *resolution* components. Thus, given a set of mentions extracted from a text document, $M = \{m_1, m_2, ..., m_n\}$ where $n$ corresponds to the number of mentions present in the text document, we aim to address the task of generating all possible coreference chains, where each coreference chain aggregates all the mentions $m_i \in M$ that refer to one and only one entity.

In the first approach to this challenging task we follow the *mention-pair model* (described in Section 2). Even aware of some limitations of this model, we argue that given the current lack of large scale and high quality annotated corpus for the task of coreference resolution in Portuguese this is the most adequate model to address this task since we can provide more training instances to the supervised machine learning classifiers. Besides this, to the best of our knowledge, the mention-pair model remains the most widely used model by the community. As a consequence, we can use current state-of-art work to compare with.

We train classifiers that learn how to classify pairs of mentions as coreferent (CO) or non-coreferent (NC), formulating the problem of coreference resolution as a binary classification problem.

The mention-pair model can be divided in three steps, namely: training set creation, mention-pair classification and coreference chains generation. In the remainder of this section, a detailed description of each of these steps is presented.

### 4.1 Training set creation

To transform the annotations available in the *Summ-it*++ corpus into a set of training instances suitable for the learning process, we need to develop methods that take into consideration the approach that will be used in the following steps of the process. Since we approach the task of coreference resolution following the mention-pair model, we need to transform annotations of coreference chains into a pairwise combination of referring expressions indicating whether the pair of referring expressions is coreferent (CO label) or not (NC label) in the original text document according to the annotation provided in the *Summ-it*++ corpus.

To generate training instances following the mention-pair model, we explored the following methods:

- *Closest antecedent:* given a mention $m_j$, a positive example $\langle m_i, m_j, CO\rangle$ is generated with $m_j$ and its closest preceding antecedent $m_i$; a set of negative examples $\langle m_k, m_j, NC\rangle$ is generated with $m_j$ and any other antecedent $m_k$ occurring between $m_i$ and $m_j$. This method of negative instance selection is further explained in more detail in Soon *et al.* [14].
- *Most confident antecedent:* given a mention $m_j$, a positive example $\langle m_i, m_j, CO\rangle$ is generated with $m_j$ and (a) its closest non-pronomial antecedent $m_i$, if $m_j$ is non-pronomial, or (b) its closest preceding antecedent $m_i$, otherwise; a set of negative examples $\langle m_k, m_j, NC\rangle$ is generated similarly to the *closest antecedent* method. This method described in more detail in Ng and Cardie [10].
- *All antecedents:* a set of positive examples $\langle m_i, m_j, CO\rangle$ is generated with any pair $m_i$ and $m_j$ annotated in the same coreference chain; a set of negative examples $\langle m_i, m_j, NC\rangle$ are generated from the rest of mention pairs.
- *Balanced:* this method is similar to the *closest antecedent* method, but instead of generating a negative example $\langle m_k, m_j, NC\rangle$ for each antecedent $m_k$ occurring between $m_i$ and $m_j$, it generates only up to two negative examples with the antecedents of $m_j$ occurring closer to the mention $m_i$ (*i.e.* $\{\langle m_{i+1}, m_j, NC\rangle, \langle m_{i+2}, mj, NC\rangle\}$). This method was conceived to generate a training set containing a balanced number of coreferent (CO) and non-coreferent (NC) training instances. By only considering as negative examples the mention pairs occurring more closely to the positive example, we provide to the classifier the most relevant training examples for the task being tackled: we aim that the classifier learns to distinguish coreferent pairs from non-coreferent pairs and, therefore, this distinction is critical for neighboring pairs;

Several experiments were made to determine the impact that different training-set creation methods have on the performance of the system proposed in this paper to address the task of coreference resolution. In Section 5 the obtained results are presented.

### 4.2 Mention-pair classification

In this paper, we formulated the problem of coreference resolution (given a set of referring expressions previously extracted from the original text document) as a binary classification problem: we aim to classify each mention-pair as coreferent (CO) or non-coreferent (NC).

In order to apply machine learning algorithms it is necessary to represent the training instances by a set of numerical features. A good set of features should represent the training instances in such a way that would make it possible for the machine learning algorithms to find patterns in the data which can be used to classify instances according to the desired target labels. Following the mention-pair model for the task of coreference resolution, the feature set should take into account that the training instances correspond to mention-pairs and, therefore, should not only represent each mention with the relevant characteristics but also the relation between the pair of mentions. Based on some of the work presented in Section 2, in particular the work of Soon *et al.* [14] and Ng and Cardie [10], we represent each mention-pair with the following set of features:

- Exact string match: if, after removing determiners, prepositions and punctuation marks, the content words of $m_i$ match the content words of $m_j$, return 1; else return 0;
- Substring match: if, after removing determiners, prepositions and punctuation marks, the intersection between the content words in $m_i$ and $m_j$ is not empty, return 1; else return 0;
- Pronoun_1: If $m_i$ is a pronoun return 1; else return 0;
- Pronoun_2: If $m_j$ is a pronoun return 1; else return 0;
- Definite_2: If $m_j$ starts with the word-lemma "o" or "de";
- Demonstrative_2: If $m_j$ starts with a demonstrative word-lemma return 1; else return 0;
- Number: If $m_i$ and $m_j$ agree in number, return 2; If they disagree, return 1; If the number information for one or both mentions cannot be determined, return 0;
- Gender: If $m_i$ and $m_j$ agree in gender, return 2; If they disagree, return 1; If the gender information for one or both mentions cannot be determined, return 0;
- Both Proper Nouns: If $m_i$ and $m_j$ are both proper nouns, return 2; If exactly one of the mentions is proper noun, return 1; else return 0;
- Appositive: If $m_i$ and $m_j$ are in an appositive relationship, return 1; else return 0;
- Number of sentences: Absolute distance between $m_i$ and $m_j$ in terms of the number of sentences;
- Number of Referring Expressions: Number of mentions between $m_i$ and $m_j$.

Based on the algorithms that obtained good results for the task of coreference resolution following the mention-pair model, as reported in Section 2, we made experiments with the following supervised machine learning algorithms: *Decision Tree*, *SVM*, *Naïve bayes*, *AdaBoost* and *Stochastic Gradient Descendent* algorithm. All the machine learning algorithms previously mentioned were implemented using the library *scikit-learn* [11] for the *Python* programming language. Then, we provide the set of features previously presented to a classifier, which learns from the the training set generated by one the methods described in Section 4.1 to classify each pair of mentions as Coreferent (CO) or Non-Coreferent (NC). Next, in Section 4.3, methods to generate the final output of the coreference resolution system, the coreference chains, are presented.

### 4.3 Coreference chains generation

In this step of the mention-pair model, the system receives as input: (a) the complete set of mentions, $M = \{m_1, m_2, ..., m_n\}$ from the annotation file and (b) the classifier trained on the training set generated by one of the methods presented in Section 4.1 and using the feature set and one of the supervised machine learning algorithms mentioned in Section 4.2. As output, in the coreference chains generation step, we aim to automatically generate the set of mentions $M$ clustered in the corresponding coreference chains, where each coreference chain represents one entity that is referred in the text by all the mentions contained in the cluster.

To generate the training instances, we processed the text from left to right and, for each active mention we generated mention-pairs from right to left. As a consequence, we must consider that a classifier trained following this approach is biased to learn how to classify each pair of mentions in this setting and, therefore, the set of instances that we aim to generate predictions from the classifier should be presented following a similar procedure (even noticing that all the classifiers described in this article are deterministic and do not take into account the sequence of labellings). Based on the approach described by Soon *et al.* [14], we employ the *closest-first strategy* to generate the coreference chains: for each mention $m_j$ a mention-pair test instance is created, in turn, with each preceding mention $m_i$ and is presented to the coreference classifier. For each pair the classifier returns the predicted label indicating whether the pair of mentions are Coreferent (CO) or Non-Coreferent (NC). This procedure terminates as soon as an antecedent is found for $m_j$ or the beginning of the text is reached. This process is repeated for each mention occurring in the text. In the end, some of the mentions are connected pairwise and, possibly, some of them are not connected to any other mention. Then, a cluster is created for each set of connected mentions (representing one coreference chain) and the mentions that are not connected to any other mention constitute one cluster (creating a coreference chain that contains only one mention). In the end of this process, we obtain the set of coreference chains predicted by the mention-pair model.

## 5 Results

The results presented in this section were obtained using the mention-pair model (described in Section 4) and using the annotations available in the *Summ-it*++ corpus (described in Section 3).

The best overall results in the baseline scenario, in which the dataset was created using the *Closest antecedent* method, were obtained with the *AdaBoost* algorithm using *Decision Trees* as weak classifiers. Therefore, we selected this supervised machine learning algorithm for all the experiments reported in this article. Furthermore, the set of features used to obtain these results remains unchanged throughout the experiments presented in this section, corresponding to the complete set of features presented in Section 4.2. Since we aim to study how different strategies used to generate training instances for the mention-pair model may influence the results that are obtained by the system, it is mandatory to keep all the remaining variables of the experimental setup unchanged throughout the experiments in order to obtain results that are scientifically valid.

| | Number of<br>CO Pairs | Number of<br>NC pairs | Avg.<br>precision | Avg.<br>recall | Avg.<br>f1-score |
|---|---|---|---|---|---|
| All antecedents | 3320 | 38759 | 0.72 | 0.20 | 0.31 |
| Closest Antecedent | 1273 | 8701 | 0.86 | 0.25 | 0.39 |
| Most Confident | 1267 | 8880 | 0.84 | 0.27 | 0.40 |
| Balanced | 2871 | 4351 | 0.64 | 0.41 | 0.50 |

**Table 1.** Mention-pair classifier CO label scores

| | Number of<br>CO Pairs | Number of<br>NC pairs | Avg.<br>precision | Avg.<br>recall | Avg.<br>f1-score |
|---|---|---|---|---|---|
| All antecedents | 3320 | 38759 | 0.94 | 0.99 | 0.96 |
| Closest Antecedent | 1273 | 8701 | 0.90 | 0.99 | 0.96 |
| Most Confident | 1267 | 8880 | 0.90 | 0.99 | 0.96 |
| Balanced | 2871 | 4351 | 0.69 | 0.85 | 0.76 |

**Table 2.** Mention-pair classifier NC label scores

The results depicted in Table 1 and Table 2 were obtained in a ten-fold cross-validation scenario. For every type of label and for the overall performance measure, the precision, recall and f1-score were used as evaluation metrics. These metrics are commonly used in classification systems to evaluate their performance.

We obtained better overall results when creating the training set using the method *All Antecedents*, which we associate with the higher number of training instances that were created with this method. However, as shown in Table 1 the dataset is very unbalanced and the performance scores for the label CO are relatively low. To overcome this problem, we made several experiments aiming to increase the performance of the algorithm for CO mention-pairs. From Table 1, we conclude that methods that create a training set with a balanced number of CO and NC mention-pairs improve the overall performance of the system in the classification of CO pairs. However, the overall performance of the system decreases, specially the classification of NC pairs (Table 2), as the number of training instances decreases.

From the results presented in this section, we conclude that increasing the size of the training set can improve the overall performance of the system and, that different methods to create training sets for the mention-pair model can have a significant impact in the quality of the predictions made by the system (i.e. performance of the system for CO mention-pairs).

From the critical analysis of the results presented by the system, we observed that exploring semantic-based features (e.g. synonyms, hyperonyms, hyponyms, word embeddings) can improve the performance of the system when the mention-pairs are noun phrases. Furthermore, we noticed that the system has low performance to detect coreference in the presence of pronouns.

In future work, we aim to improve the results presented in this section exploring semantic-based features and, addressing the problem of pronominalization.

## 6 Conclusions

In this article, we described a preliminary approach to address the problem of coreference resolution from text written in the Portuguese language. We addressed this problem using the mention-pair model and exploring features at the lexical, morphological, syntactic and structural level. In addition, we explored different methods to generate the training set from an annotated corpus. The critical analysis of the experimental results presented in this article demonstrates the difficulty of the task, particularly when the corpus used in our experiments contains a relatively low number of annotations. In future work, we expect to improve the results presented in this article using a more sophisticated set of features (for instance, semantic-level features), addressing the problem of pronominalization and, exploring different approaches to address the task of coreference resolution in Portuguese (e.g. entity-based models).

**Acknowledgments.** This work was partially supported by a doctoral scholarship provided by the Doctoral Program in Informatics Engineering (ProDEI) from the Faculty of Engineering of the University of Porto (FEUP).

## References

1. Coreixas, T.: Resolução de correferência e categorias de entidades nomeadas. Master's thesis, Pontifícia Universidade Católica do Rio Grande do Sul (2010)
2. Denis, P., Baldridge, J.: Specialized models and ranking for coreference resolution. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. pp. 660–669. EMNLP '08, Stroudsburg, PA, USA (2008)
3. Fonseca, E.B., Antonitsch, A., Collovini, S., do Amaral, D.O.F., Vieira, R., Figueira, A.: Summ-it++: an enriched version of the summ-it corpus. In: Proceedings of the Tenth International Conference on Language Resources and Evaluation LREC 2016, Portorož, Slovenia, May 23-28, 2016. (2016)
4. Fonseca, E.B., Vieira, R., Vanin, A.A.: Improving coreference resolution with semantic knowledge. In: Silva, J.R., Ribeiro, R., Quaresma, P., Adami, A., Branco, A. (eds.) Computational Processing of the Portuguese Language - 12th International Conference, PROPOR 2016, Tomar, Portugal, July 13-15, 2016, Proceedings. Lecture Notes in Computer Science, vol. 9727, pp. 213–224. Springer (2016)
5. Freitas, C., Mota, C., Santos, D., Oliveira, H.G., Carvalho, P.: Second harem: Advancing the state of the art of named entity recognition in portuguese. In: Chair), N.C.C., Choukri, K., Maegaard, B., Mariani, J., Odijk, J., Piperidis, S., Rosner, M., Tapias, D. (eds.) Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10). European Language Resources Association (ELRA), Valletta, Malta (may 2010)
6. García, M., Gamallo, P.: An entity-centric coreference resolution system for person entities with rich linguistic information. In: COLING 2014, 25th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, August 23-29, 2014, Dublin, Ireland. pp. 741–752 (2014)
7. Garcia, M., Gamallo, P.: Multilingual corpora with coreferential annotation of person entities. In: The 9th edition of the Language Resources and Evaluation Conference. pp. 3229–3233. European Language Resources Association (2014)

8. Grosz, B.J., Joshi, A.K., Weinstein, S.: Providing a unified account of definite noun phrases in discourse. In: Proceedings of the 21st Annual Meeting on Association for Computational Linguistics. pp. 44–50. ACL '83, Stroudsburg, PA, USA (1983)
9. Jurafsky, D., Martin, J.H.: Speech and Language Processing (2Nd Edition). Prentice-Hall, Inc., Upper Saddle River, NJ, USA (2009)
10. Ng, V., Cardie, C.: Improving machine learning approaches to coreference resolution. In: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics. pp. 104–111. ACL '02, Stroudsburg, PA, USA (2002)
11. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: Machine learning in Python. Journal of Machine Learning Research 12, 2825–2830 (2011)
12. Rich, E., LuperFoy, S.: An architecture for anaphora resolution. In: Proceedings of the Second Conference on Applied Natural Language Processing. pp. 18–24. ANLC '88, Association for Computational Linguistics, Stroudsburg, PA, USA (1988)
13. Sapena, E., Padró, L., Turmo, J.: A constraint-based hypergraph partitioning approach to coreference resolution. Computational Linguistics 39(4), 847–884 (2013)
14. Soon, W.M., Ng, H.T., Lim, D.C.Y.: A machine learning approach to coreference resolution of noun phrases. Comput. Linguist. 27(4), 521–544 (Dec 2001)
15. Walker, M., Joshi, A., Prince, E.: Centering Theory in Discourse. Clarendon Press (1998)

# Comparison of Classification and Regression Approaches for Predicting Age of Onset

Maria Pedroto

Faculty of Engineering, University of Porto,
Rua Dr. Roberto Frias, s/n 4200-465 Porto, Portugal
up200005203@fe.up.pt

**Abstract.** This work describes a methodological approach to predict the Age of Onset of Patients diagnosed with Transthyretin Familial Amyloid Polyneuropathy (TTR-FAP). The learning methodology consisted of a supervised learning model, with a 10-fold cross validation method. The modelling approaches used were Classification and Regression. The traditional algorithms used with each of these approaches were Decision Trees, SVM and Random Forests. The results achieved show that for the Classification approach there is a better accuracy in the Decision Trees with 85% of accurately classified records. In the case of the Regression Models we achieved better results with Random Forests, with an absolute error of 3 years on average. This work is important as it can assist in reframing of the appointment scheduling process as we can use the models with better results to produce predictions for the Age of Onset of Asymptomatic Patients. These values will in due time be evaluated by the medical professionals.

**Keywords:** Supervised Learning Methodologies, Classification Models, Regression Models, SVM, Decision Trees, Random Forests, Integration.

## 1 Introduction

Amyloid Neuropathies are severe, progressive and life-threatening diseases that are characterized by sensory, motor and autonomic neuropathies [1]. Neuropathy is generally a disease that damages or affects a patient's nerves, thus impairing movement, organ function and heat sensation.

Clinically, TTR-FAP manifests itself as a progressive and irreversible disease that impairs sensory and autonomic movements [2]. A positive patient is categorized into one of three stages [2]:

- I (sensory polyneuropathy): where the patients suffer from sensory impairment, which starts of the lower members and progresses to every member of the body;

- II (progressive walking disability): when patients start showing difficulties in walking and start using crunches or a second person to assist them;
- III (wheelchair bound or bedridden): when patients lose all their ability of walking and are confined to a wheelchair or to a bed.

First cases of this disease were described in Northern Portugal, in 1952, [2] by Professor Dr. Mário Corino de Andrade. Since then several cases were diagnosed in: Sweden, France, Italy, Spain/ Majorca, Bulgaria, Germany, Netherlands, Cyprus and Turkey [2], with Portugal having the largest focus of Patients in the world with the ValMet30 genetic variant.

### 1.1. The importance of predicting the Age of Onset

Age of onset is a medical term referring to the age on which a patient first acquires, develops or experiments the first symptoms of a disease. A factor that can greatly impact the evolution of the disease is related with a "as soon as possible" effective diagnosis of the disease. So, in the cases when the patients arrive at a diagnosis center prior to the start of the first symptoms it is highly important to define models that can accurately define their expected age of onset.

### 1.2. Work Contribution and Paper Organization

Our work compares typical Regression approaches with a transformed Classification view of a medical problem. By transformed we mean that the target variable is preprocessed to represent interval values (in our results, each class represents 4 years of patients age of onset). Besides this approach which we consider to be a valuable scientific contribution, our work also represents a step forward by the systematic generation of different features engineered by the integration of familiar network data with specific patients' events. This type of data, as far as we know it, isn't referenced in any work that gives an insight as to the age of onset of specific diseases. The most approximate reference that we could find is in [3] where the authors use several features engineered from hierarchical datasets to reveal lifespan patterns in human population.
This paper is organized as follows: in section 1 we define an introduction to the field of research; in section 2 we define our feature engineering process and give some insights into the dataset used; in section 3 we define our experimental working environment, the experiments made and the results; in section 4 we define previous work related with this paper; in section 5 we state our main conclusions related with this work; and in section 6 we define the guidelines for future work in this field.

## 2 Data Analysis and Feature Engineering

### 2.1. Dataset

This work is based on datasets created by the integration of disparate sources of information managed by Unidade Doutor Corino de Andrade (UCA), of Centro Hospitalar do Porto (Chporto). These files are: Patients Hierarchical Family Structure, Patients Personal Information and Patients Exams Information. The Patients Hierarchical Family Structure was compiled by exporting Cyrillic datasets into Gedcom format, which is a specific Genealogy data exchange format [4]. Cyrillic is a program used to design a patients' ancestral pedigree. These files concentrate a great deal of information about a patients' familiar structure (date of birth of patients' familiar relations; average dimension of familiar structure) that isn't easily available in traditional patient's electronic data records (EDR). After the exporting process, we created an ETL framework that: retrieves the information in each file and cleans, transforms and normalizes the fields necessary to be processed and loaded into the analytical SQLite database. Later, the same SQLite database integrates the Patients Personal Information extracted from the respective Microsoft Access Database. This database, migrated from early dBase files, contains mainly management information (patient's id, date of insertion in the system, date of first appointment) and specific events related with the patients. In the last step of the process we extract, transform and load the information related with the different Patients Exams Information Database. This file contains records related with the exams performed to the patients during the follow up appointments.

### 2.2. Data Analysis (Patient Data)

In **Table 1** we have an aggregated view into the distribution of Symptomatic and Asymptomatic patients diagnosed at UCA. From the results we can infer that almost every patient with the disease has died in the Symptomatic stage. This is due to the fact that most patients that are systematically followed in a medical facility have already passed by the first symptoms phase.

**Table 1.** Total Distribution of Patients' Status (patients diagnosed in Unidade Corino de Andrade and who allowed for their information to be harvested)

| **Symptoms** | **Sex** | **Patients** | |
|---|---|---|---|
| | | **Alive** | **Dead** |
| Symptomatic (2846) | Male (1533) | 41.3% (633) | 58.7% (900) |
| | Female (1313) | 40% (526) | 60% (787) |
| Asymptomatic or Pre-Symptomatic (681) | Male (238) | 95% (225) | 5% (13) |
| | Female (443) | 97% (430) | 3% (13) |

As stated before, when profiling the records, we took into consideration specific life events data. These events are: birth year, death year and patient's onset year (see **Fig. 1**). With them we calculated the variables for: age at onset; current age (for patients that haven't died yet); and age at death. With the distribution plots we state that: most patients have an age of onset between 30 and 40 years old and they were diagnosed between 1990 and 2010; most of the patients are between 40 and 60 years old, and were born between 1950 and 1970; finally, that most of the patients died around 40 and 60 years old around 2000 and 2010.

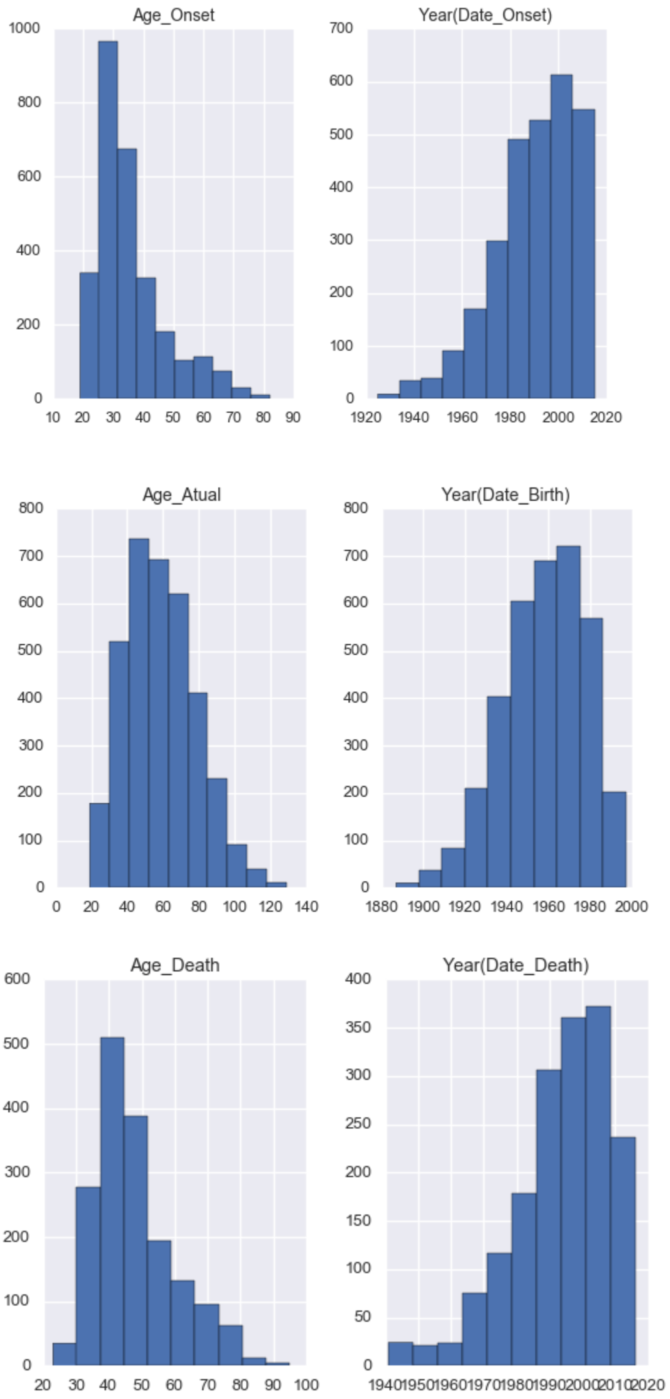


**Fig. 1.** Patients' Life Events Data Distribution

A main aspect visible by the distribution plots relates with the general distribution of these variables: they are highly skewed and unbalanced.

### 2.3. Summary Statistics Analysis

Usually, summary statistics help us get a closer look at specific variables distribution. With the summary statistics available in **Table 2** it is visible that: the youngest patient in the dataset is 19 years old; the amplitude for the patients' year of death goes from 1941 into 2016; the patients were born over 110 years (from 1887 to 1997); the first patient was diagnosed with symptoms in 1925; the amplitude for the year of the age of onset is of 90 years (very high number of different values); the date of death has 1814 missing, unknown or not yet reached values (patients' that have not died yet); the year of age of onset variable has 707 missing, unknown or not yet reached values (these patients are known as Asymptomatic Positive Carriers); the youngest patient died when he was 24 years old and the oldest died with 91 years old.

**Table 2.** Summary Statistics for Patients' Continuous Data

| **Statistic** | **Year(Date_Birth)** | **Year(Date_Death)** | **Year(Age_Onset)** |
|---|---|---|---|
| **Count** | 3527 | 1713 | 2820 |
| **Average** | 1957.5 | 1993.27 | 1999.72 |
| **SD** | 19.7 | 14.78 | 16.8 |
| **Minimum** | 1887 | 1941 | 1925 |
| **25%** | 1944.5 | 1985 | 1980 |
| **50%** | 1959 | 1996 | 1992 |
| **75%** | 1973 | 2004 | 2003 |
| **Maximum** | 1997 | 2016 | 2015 |
| **Statistic** | **Age_Actual** | **Age_Death** | **Age_Onset** |
| **Count** | 3527 | 1713 | 2830 |
| **Average** | 58.5 | 47.8 | 35.7 |
| **SD** | 19.7 | 12.12 | 11.4 |
| **Minimum** | 19 | 23 | 19 |
| **25%** | 43 | 39 | 28 |
| **50%** | 57 | 45 | 32 |
| **75%** | 71.5 | 54 | 40 |
| **Maximum** | 129 | 95 | 82 |

### 2.4. Feature Engineering

For our methodological approach, and considering the information that we can extract from the familiar Cyrillic trees, we decided to generate three types of features regarding TTR-FAP Patients: Patient Dataset; First Level Family Dataset; and Expanded Family Dataset.

For the Patient Dataset (see **Table 3**) we maintained the variables analyzed before, where we focused on specific life events of the patients. This is due to their impact in reflecting the main aspects in a patients' life. After processing this information, we evaluated the type of information that we could extract from a First Level Familiar Structure and the respective Family Network Files (**Table 4**). In this case, we considered the patient's Siblings and Children. So, we calculated: the number of siblings and descendants for either sex, with and without the disease; the average, maximum and minimum age at onset of children and siblings, also of either sex; and the average, maximum and minimum difference of birth dates for the Patients Siblings.

**Table 3.** Patients' Features Dataset

| Feature | Description |
|---|---|
| Sex | The sex of the Patient (1 for males and 2 for females) |
| Year of Birth | Year of birth of the Patient |
| Year of Death | Year of death of the Patient |
| Year of Onset | Year of Patients' First Symptoms |
| Actual Age | Patients' Current Age |
| Age of Death | Patients' Age of Death |
| Age of Onset | Patients' Age When Diseased |
| Year of Onset of Parent | Year of Onset of Parent |
| Age of Onset of Parent | Age of Onset of Parent |

**Table 4.** First Level Family Features Dataset

| Feature | Description |
|---|---|
| Number of Sons and Daughters | Number of Descendants (by sex) |
| Number of Brothers and Sisters | Number of Siblings (by sex) |
| Number of Sons and Daughters with the Disease | Number of Descendants with the disease (by sex) |
| Number of Brothers and Sisters with the Disease | Number of Siblings with Disease (by sex) |
| Avg, Max and Min Age at Onset of Brothers and Sisters | Average, Maximum and Minimum Age at Onset of Siblings (by sex) |
| Avg, Max and Min Age at Onset of Sons and Daughters | Average, Maximum and Minimum Age at Onset of Descendants (by sex) |
| Avg, Max and Min difference Years of Birth of Brothers and Sisters | Average, Maximum and Minimum difference between years of Birth Descendants (by sex) |
| Avg, Max and Min difference Years of Birth of Descendants | Average, Maximum and Minimum difference between years of Birth of Descendants (by sex) |

The features related with the Family Network Files are used to account for specific correlation between Extended Familiar Families, as some families, for example, have a higher number of generations then others (see **Table 5**). In this case, we used: the total number of patients with the disease, their age of onset difference and the average, maximum and minimum age of onset of their respective patients.

### 2.5. Data Quality

During any knowledge data discovery research activity, an important step relates with the definition of metrics to help to effectively check for data consistency. In **Table 6** we identify the main types of problems that were dwelt with during the integration phase. Also, during this phase, and later during the profiling of the dataset and evaluation of the features related with the difference in years of birth of First Level Families, we found an interesting event. Some of the values generated had to be wrong as, for example, the difference in years between parents and their descendants has small and specific invalid outliers. We state these values as outliers, and as naturally invalid as it is impossible to have a difference of age between a parent and his children smaller than 8 years. To deal with this problem we defined a rule to validate data where if this difference is below 8 we update the value to 0. We based ourselves on [5], where the authors identify the minimum and maximum age of puberty for American Children. There, 8 is referenced as the minimum age of puberty for female American Children.

**Table 5.** Extended Tree Features

| Feature | Description |
|---|---|
| Avg, Min and Max Age of Onset of Patients belonging to each Tree | Average, Minimum and Maximum Age of Onset of Symptomatic Patients |
| Difference between the Maximum and Minimum Age of Onset | Difference between the Maximum and the Minimum Age of Onset |
| Count Patients | Number of Positive Patients |

**Table 6.** Main Data Quality Problems

| System | Type of Inconsistency |
|---|---|
| Microsoft Access Databases | Wrong sex specification |
| | Incomplete fields |
| Cyrillic Hierarchical Structures | Wrong sex specification |
| | Inconsistent Name |
| | Wrong Year of Birth or Year of Death |
| | Wrong Patient ID reference |
| | Invalid reference to dates between hierarchically referenced individuals |

# 3 Methodology, Experimental Context, and Evaluation

In this work, we used Decision Trees, Support Vector Machines (SVM) and Random Forests to define supervised learning models to assess for the Age at Onset of Symptomatic Patients with TTR-FAP. This is a supervised learning approach as we will only be using data which has the respective values for age of onset. The modelling approaches are regression and classification based. In the next sections, we will be: defining the parameters used; identifying the programming packages and respective programming language version; explaining the modelling approaches considered; and comparing the different models with a set of different error metrics for error evaluation.

## 3.1. Developing Environment

For this work, we used pipelines constructed in Python version 3.5.2. We used Conda as a development environment to support the usage of Jupyter Notebooks. Also, the main libraries used were: scikit-learn (version 0.18.1); seaborn (version 0.7.1); graphviz (version 2.38.0); and pandas (version 0.19.0).

## 3.2. Modelling Techniques

The modelling approach used for all the different algorithms approaches is the same, may it be for the Classification or Regression models. In **Fig. 2** it is possible to check the entire process that is applied to each tuple of the train and test sets.

The process starts with the extraction of the Records related with the Symptomatic Patients. Due to unbalance problems, we used Oversampling to work with the data before splitting into 10-Fold buckets. This technique is a well-known probabilistic approach to correct unbalanced datasets. There are other techniques available in Python and related with undersampling and hybrid approaches. These techniques weren't considered for this work due to its low amount of data. For these models, we won't be using the "Year (Age at Onset)" variable as it has a high correlation with the label variable that we are trying to predict (in all our experiments our target variable in the Age of Onset). In the case of the Classification Approaches we will be considering bins of 4 ages length in a total of 20 classes.

## 3.3. Experiments

In **Table 7** and **Table 8** we have the results of applying Classification and Regression Models to several different types of algorithms. There are well known classical approaches used for supervised learning.

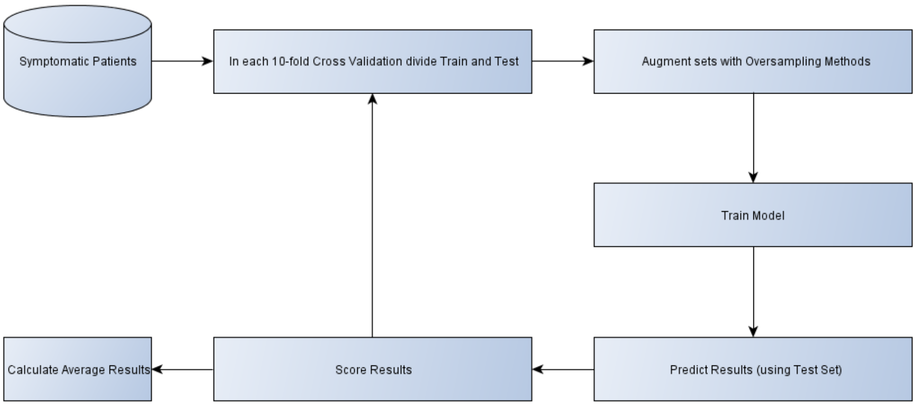

Symptomatic Patients
In each 10-fold Cross Validation divide Train and Test
Augment sets with Oversampling Methods
Train Model
Predict Results (using Test Set)
Score Results
Calculate Average Results

**Fig. 2.** 10-Fold Cross Validation Process Modeling

In the case of the Decision Tree Classification Experiments we used, in approach A, as models parameters: a criterion of 'entropy', a tree maximum depth of 10, a minimum of value for sample split equal to 4 and minimum value for samples in leaf of 4. Also, in the case of approach B, we used as models parameters: a criterion of 'entropy', a tree maximum depth of 20, a minimum value for sample split of 8 and a minimum sample in leaf equal to 8. In the case of SVM Classifiers we used the original parameters set by scikit-learn framework. For the Random Forests, we used the same parameters as those applied for the Decision Trees, and we used 100 different trees. In all the approaches we considered a factor of 4 years for each different class to be predict.

**Table 7.** Results for comparison of Classification Approaches using Decision Trees, SVM and Random Forest

| **Algorithm** | **Metric** | **Classifier (A)** | **Classifier (B)** |
|---|---|---|---|
| Decision Tree | Precision | **0.85** | 0.82 |
| | Recall | **0.86** | 0.83 |
| | F1-score | **0.86** | 0.83 |
| SVM | Precision | 0.25 | |
| | Recall | 0.1 | |
| | F1-score | 0.09 | |
| Random Forest | Precision | 0.81 | 0.80 |
| | Recall | 0.82 | 0.80 |
| | F1-score | 0.81 | 0.79 |

In the case of the Regression Tree Classification Experiments we also used two different approaches. In the case of approach A we have: a criterion of mean squared error, a maximum depth of 10, a minimum of sample split equal to 4 and a minimum sample value in the leaves of 4. In the case of approach B, we have: a criterion of mean squared error, a tree maximum depth of 20, a minimum value for sample split of 8 and a minimum sample in leaf equal to 8. For the Support Vector Machine algorithm, we maintained the original parameters setup provided by scikit-learn framework. When it came to the Random Forests algorithm, we used the same values as those referred in the Decision Tree, besides using an estimator of 100. So, in total, we conducted 10 experiments.

**Table 8.** Results for comparison of Regression Approaches using Decision Trees, SVM and Random Forest

| Algorithm | Metric | Regression (A) | Regression (B) |
|---|---|---|---|
| Decision Tree | MAE | 3.5 | 3.7 |
| | MSE | 26.9 | 31.9 |
| | RMSE | 1.6 | 1.8 |
| SVM | MAE | 13.3 | |
| | MSE | 234.6 | |
| | RMSE | 4.8 | |
| Random Forest | MAE | 3.32 | **3.1** |
| | MSE | 24.6 | **22.7** |
| | RMSE | 1.57 | **1.5** |

## 4 Related Work

### 4.1. Statistical Approaches

To accurately predict the Age at Onset of Neurological Diseases there are specific statistical approaches that have been widely used over the past decades. The most common deals with the usage of Linear Regression Models as these datasets generally have a common tendency among their records. This can be checked, for example, in [6] where the authors use linear regression algorithms to estimate the age at onset of a patient due to parental factors. This procedure was also used by [7].

### 4.2. Data Mining and Machine Learning Approaches in Medical Sciences

During the past few years several different Data Mining and Machine Learning Approaches have been used to advance the field of Medical Sciences. The scientific community has seen great innovations coming from this field. Ever since the usage of Machine Learning techniques to generate new chemical combinations to treat certain diseases, to the usage of Text Mining approaches to extract knowledge from clinical records, without forgetting the most recent usage of Machine Learning approaches to support clinical diagnosis it seems that this is a thriving and active field area.

Several of the recent works in the field of Medical Data Sciences use network approaches to assist in the resolution of different medical problems that exist in world. For example, in [8], the authors used a phenotypic comorbidity network to understand

the structural connections between diseases, while in [9] the authors use familial relations to predict the risk of a future patient to acquire cancer.
Although not related with the application of Data Mining and Machine Learning approaches in the field of medical sciences in [3] the authors use a high number of data mining algorithms (OneR, Decision Tree, Naïve-Bayes, Knn, Random Forest, Bagging) to extract knowledge from genealogical trees available online. This is a very valuable work as they show the type of features that can be engineered from genealogy trees.

## 5 Conclusions

The usage of Data Mining and Machine Learning approaches for helping to ensure a better medical assistance to patients with specific medical conditions is starting to become an accepted approach in most medical centers. This work defines a modelling approach that can help to produce insights into the reuse of information being produced in these facilities. However, the quality of the results can be greatly impacted by the quality of the data being collected. We believe that the integration of disparate sources of information can help to improve the quality of these valuable assets.

In our work, we used the distribution of onsets of the patients and their familiar network. We believe that this approach can help ensure acceptable results in the cases where patients' ancestors aren't being monitored by the same hospital facility or when they have already passed away. Also, in the cases where there is only information about the Symptomatic Phase of the Parents, this approach can give some orientation guidelines. To help treatment of these cases our approach combines the results of a patient's family to give some indications about the near future.

Our results show that Decision Trees produce the best results in the Classification approach A (with an average of 85% of correct values classified). In the Regression models it is clear that the most accurate model is by far Regression Models approach B with an average of 3 years for absolute error. Althought these aren't specially good results, we believe that with further work the values can be improved. In the case of the SVM results we acknowledge that the values for the Classification and Regression Approaches are high. This fact will be target of further work in the near future as we believe that in part this can be due to the usage of the default parameters of the scikit-learn, and by the not normalization and regularization of the features in the pre-processing stage.

In the next section we present some guidelines for the next steps to be conducted.

## 6 Future Work

This work can be expanded with an extended study of the different algorithms available at the scikit-learn framework. Some of these algorithms represent specific multiclass approaches that should be better suited for our problem and could help improve the accuracy of the results presented. Also, the modelling approaches can be redefined so as to acknowledge a different weight for prevision errors in the cases where the prevision is bigger or smaller than the actual one (Classification Approaches). Although we are considering interval classes of labels of size 4, we believe that further work can be done in analyzing the sensibility of this factor. A more methodological evaluation of the feature engineering process can help improve the quality of the models. Other models may also be developed by the integration of different sources of information (for example patients exams). Different approaches of working with missing values can also be performed and evaluated (we are naively inputting the value of 0).

**Acknowledgments.** We would like to thank the valuable contribution of Professor Alípio Jorge and Professor João Mendes Moreira of Laboratory of Artificial Intelligence and Decision Support (LIAAD) of INESC TEC; of Dra. Teresa Coelho from Hospital de Santo António do Porto; and Paulo Marques from the Faculty of Sciences of University of Porto (FCUP). We would also like to thank Centro Hospitalar do Porto for supporting this work.

## References

1. Y. Ando, T. Coelho e J. Berk, "Guideline of transthyretin-related hereditary amyloidosis for clinicians," *Orphanet Journal of Rare Diseases,* pp. 8-31, 2013.
2. Y. Parman, D. Adams, L. Obici, L. Galán, V. Guergueltcheva e O. B. Suhr, "Sixty years of transthyretin familial amyloid polyneuropathy (TTR-FAP) in Europe: where are we now? A European network approach to defining the epidemiology and management patterns for TTR-FAP.," *Current Opinion in Neurology, 29(Suppl 1), S3–S13,* February 2016.
3. M. Fire e Y. Elovici, "Data Mining of Online Genealogy Datasets for Revealing Lifespan Patterns in Human Population," 2013.
4. "wikitree," 15 5 2016. [Online]. Available: https://www.wikitree.com/wiki/GEDCOM.
5. J. A. H. Care, "Normal ages of pubertal events among American males and females," *Journal of Adolescent Health Care,* vol. 1, pp. 26-29, 1980.
6. C. Brackenridge e B. Teltscher, "Estimation of the age at onset of Huntington's disease from factors associated with the affected parent," *Journal of Medical Genetics,* nº Medical Genetics, pp. 64-69, 1975.
7. M. Orth e C. Schwenke, "Age-at-onset in Huntington Disease," , 2011.
8. F. Folino, C. Pizzuti e M. Ventura, "A Comorbity Network Approach to Predict Disease Risk".
9. R. Chau e A. K. Win, "Determining the familial risk distribution of colorectal cancer: a data mining approach," *Familial Cancer,* 2016.
10. T. Coelho, I. Conceição, M. Neves e A. M. Silva, "Evaluation of one year treatment with Tafamidis in Portuguese Patients with Transthyretin Familial Amyloid Polyneuropathy," em *Journal of Neurology,* 2014.

# Combining Forecasters using Arbitration for Water Consumption Forecasting

Vítor Cerqueira[1] and Luís Torgo[2]

[1] FEUP - University of Porto, LIAAD-INESC TEC, Porto, Portugal
cerqueira.vitormanuel@gmail.com,
[2] LIAAD-INESC TEC, Porto, Portugal

**Abstract.** Water utilities systems rely on water consumption forecasting models to plan for future operations. They use predictive methods to control reservoir and distribution pressure levels while minimizing costs and maximizing water quality. This paper proposes a novel method for water consumption forecasting. The method is based on an ensemble approach, in which individual competing models are arbitrated by a metalearning layer. The goal of arbitrating individual forecasters is to dynamically combine them according to their aptitude in the input data. We validate our proposed model for short-term water consumption forecasting using a real-world water provider. The results from empirical experiments show that the proposed method is competitive with other methods, including current state-of-the-art methods used for time series forecasting tasks.

**Keywords:** Ensemble Methods, Short Term Water Consumption Forecasting, Metalearning, Time Series, Bias Weighting

## 1 Introduction

Water utility systems use short-term water consumption – henceforth denoted as STWC – forecasting techniques to plan their operations in an efficient manner. The optimization of the pumping schedule and water treatment strategies enables operation planners to reduce energy and treatment costs while maximizing the quality of supplied water.

The optimization of operations in water utilities is economically essential. For example, San Diego Water Department achieved savings of approximately $800.000 US dollars on their first year after introducing a water consumption forecasting system in their short-term planning [10]. In this study we focus on STWC forecasting, predicting demand levels in a granularity of day. In this context, these daily projections of water consumption support decisions concerning the short-term management of the water supply grid. This paper presents an novel time series model for STCW forecasting. The proposed method leverages the predictive power of ensemble methods, combining individual learning models with different inductive bias using a metalearning strategy. We explore ways of combining the predictions of forecasters in a dynamic – online – fashion. In time evolving environments the process causing the underlying data is prone to change over time and the combined model should adapt accordingly.

We use metalearning (e.g. [4]) to analyze the expertise each individual forecaster across the time series of water consumption. We can then use this meta knowledge to dynamically weight the predictions of base learners according to their competence in a given observation. If we expect a given forecaster to perform poorly in some subset of the data we render it a low weight in the combination rule. On the other hand, if we are optimistic about some learner in our ensemble we augment its weight with respect to the other learners.

The intuition behind our approach is that different learning models may have different areas of expertise across the input space. That is, in a given test observation, some individual learner may be better than the combined model and different individual learners will be better on different cases. In effect, we can learn about the learning process of each base-learner. In fact, we hypothesize that the underlying process causing the series of water consumption follows a recurring pattern due to seasonal factors. Consequently, we hypothesize that the metalearning layer enables the combined model to better detect changes between different regimes and quickly adapt itself to the environment.

Our Metalearning strategy follows an Arbitrating scheme ([21] and [13]), in which we build a meta-learner for each base-learner comprising the ensemble. Each meta-learner is specifically designed to assess how apt its base counterpart is to make a prediction in a given observation. This is accomplished by analyzing how the error incurred by a given learning model relates to the characteristics of the data. The meta-features (feature set in the meta level – e.g. [23]) used to build the meta-learners are the primitive features used to learn the target concept along with some summary statistics which describe the recent dynamics of the data. At test time, the base-learners are weighted according to their degree of competence in the input observation, estimated by the predictions of the meta-learners. Specifically, a meta-learner predicts the loss (e.g. absolute error) that its base-level associate will incur in some observation.

Our goal is to predict the next value of the series of water consumption. We use regression models as our base-learners by transforming the water consumption time series into an Euclidean space using time delay embedding [27]. Furthermore, in order to augment the information about the data we also use external predictors such as weather reports. In summary, the contributions of this paper are the following:

- An arbitrated ensemble for STCW forecasting. The ensemble includes a meta-learning layer based on an arbitrating scheme, used to dynamically combine individual models;
- We use the Arbitrating strategy to dynamically weight individual models, whereas typical applications select the most reliable model at each test query.

We start by outlining the related work in Section 2; the methodology is addressed in Section 3, where we formalize and explain our contributions; the Case Study is briefly described in Section 4, along with the pre-processing steps and descriptive statistics. The experiments and respective results are presented and discussed in Section 5. Finally, the conclusions are produced in Section 6, along with some remarks about future work and reproducibility of the proposed methods.

## 2 Related Work

In this paper we focus on ensembles with self-adapting mechanisms to predict STWC, which is a time series with a numerical outcome. Ensemble methods for numerical predictions problems have a vast literature. We refer to the survey written in [18] for a complete overview on ensemble approaches for these tasks.

Building adaptable models is important in dynamic real-world environments in which data is constantly changing over time due to several factors, for example, seasonalities. Our proposed method is motivated by the core concepts behind Arbitrating classifiers ([13], [21]). Arbitrating is an ensemble method used to combine classifiers according to their expertise on the input data. The expertise of a base-learner is learned with a corresponding meta-model, which learns the loss of its base counterpart according to a set of meta-features. At test time, the classifier with greatest confidence in the input data-point is selected to make a prediction. The authors reason that each meta-model holds information about which part of the data its base counterpart works best and considers when it can make a reliable prediction.

Other forms for dynamically combining models for time series forecasting with numerical outcome were proposed in the literature. In [24], the authors use Zoomed Ranking [26] approach to rank and select time series forecasting models. MetaStream is proposed in [23]. The authors summarise the dynamics of the recent and upcoming observations in a data stream to either select or combine regression models. In [29] and [15] presented other two approaches that use the characteristics of time series in a meta-level to improve the combination of individual forecasters. They use these characteristics to induce several rules to weight or select between different models.

Our approach is different from the existing literature in the sense that we apply an arbitrating scheme to meta-learn and weight the individual base-learners. To the best of our knowledge, this is the first application of an Arbitrating scheme for time series prediction with numerical outcome, particularly STWC forecasting.

### 2.1 Water Consumption Forecasting

Long-Term Water Consumption Forecasting typically covers a time span from 10 to 20 years. It essentially relates to the system planning of expanding or construction of new facilities and pipelines [10]. In this paper we focus on STWC Forecasting. Water utilities typically use daily forecasts to optimize system operations regarding water delivery, water treatment and distribution system pumping.

Several STWC forecasting models have been proposed in the literature. The most typical approaches rely on regression and time series analysis models (e.g. [17], [16], [25], [32]). The connectionist approach of Artificial Neural Networks is also of common use, for example in [9], [33] or [7]. The temporal granularity in these approaches ranges from hourly to monthly forecasts.

Many approaches also incorporate external features in their methodology, such as [2] or [3]. These typically include weather information such as temperature and rainfall indicators.

## 3 Short-Term Water Consumption Forecasting

Formally, water consumption forecasting is a particular problem of time series forecasting tasks. We start addressing the methodology by presenting the main notation employed throughout this section:

**Time Series:** A time series is a temporal sequence of values $Y = \{y_1, y_2, \dots, y_t\}$, where $y_i$ is the value of $Y$ at time $i$ and $t$ is the length of $Y$;

**Embedded Time Series:** $Y^K$ denotes the embedded time series with embedding dimension $K$. We use time delay embedding to represent $Y$ in an Euclidean space with embedding dimension $K$, according to [27]. In effect, we generate the following matrix:

$$Y^K = \begin{bmatrix} y_1 & y_2 & \cdots & y_{K-1} & y_K \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ y_{i-K+1} & y_{i-K+2} & \cdots & y_{i-1} & y_i \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ y_{t-K+1} & y_{t-K+2} & \cdots & y_{t-1} & y_t \end{bmatrix} \tag{1}$$

Each row denotes an embedding vector $v_r$, $\forall\ r \in \{1, \dots, t-K+1\}$. Our goal is to predict the next point in the series, represented by the last column in Matrix 1;

**External Predictors:** $Y^{ext}$ denotes the set of external predictors computed for each embedding vector $v \in V$. These include external information (e.g. weather data) which helps to model the target concept;

**Base-Learners:** We denote as $M$ the set of $m$ base-learners comprising the ensemble $S$;

**Meta-Learners:** $\overline{M}^j$ is a meta-learner for $M^j$, with $j \in \{1, ..., m\}$;

**Base-Learners Loss:** $e_i^j$ represents the absolute loss of $M^j$ in the observation $y_i$;

**Base-Learners weights:** $w_i^j$ denotes the weights assigned to $M^j$ for predicting the value of $y_i$.

Our methodology for STWC forecasting settles on the three main steps: An offline (**i**) training step of $M$ and the online iterative steps: (**ii**) Meta-learning of $\overline{M}$ and (**iii**) prediction of $y_{t+1}$ using $M$ which is dynamically weighted according to $\overline{M}$.

### 3.1 Learning $M$

In the first step we train the learning models $M$ which are then combined to make a prediction. Concretely, each $M^j, \forall\ j \in \{1, ..., m\}$ is individually trained using the available $Y_{tr}^K$, the embedded time series combined with the external predictors $Y^{ext}$. $M$ is composed of individual regression models with different inductive bias. Different models (e.g. Gaussian Processes and Neural Networks) hold different assumptions regarding the underlying data. This divergence across base-learners comprising $S$ encourages diversity in the ensemble – a fundamental ingredient in the ensemble recipe [5].

### 3.2 Metalearning $\overline{M}$

The metalearning step of our methodology is an online process run at test time. Our objective in applying this metalearning strategy is to extract the information about the expertise of each individual model in $M$ across the series of water consumption.

We use a metalearning layer for arbitrating among competing individual learners. However, instead of selecting the most reliable model (as in [13] and [21]), we use the meta-knowledge to weight the base learners according to their expertise in the input signal.

Formally, each meta-learner $\overline{M}^j, \forall\ j \in \{1, \ldots, m\}$, is trained to build a model for $e^j = f(\overline{X})$, where $f$ denotes the regression function. $\overline{X}$ represent the meta-features, i.e., the set of **features** used in the **meta**-level by the meta-learners in $\overline{M}$. $\overline{X}$ is composed by the primitive features used by $M$ along with some summary statistics. These statistics are computed for each each embedding vector and characterize the recent dynamics of the series as well as its structure.

We conduct this meta regression analysis to understand how the loss of a given base-learner relates to the different dynamics of the series. In effect, we can explore forms of capitalizing from these relationships. Specifically, we use the information from $e$ to dynamically weight the base-learners $M$.

### 3.3 Predicting $y_{t+1}$

When a new observation $y_{t+1}$ arrives for prediction we combine the predictions of $M$ with the meta information from $\overline{M}$. The arbitrating layer composed by $\overline{M}$ is able to predict how well each base learner in $M$ will perform with respect to each other. If $\overline{M}^j$ predicts that its counterpart $M^j$ will make a large error ($\hat{e}^j_{t+1}$) relative to the other base learners ($\hat{e}^l_{t+1}, \forall\ l \in \{1, \ldots, m\} \backslash \{j\}$) then $M^j$ will be assigned a small relative weight in the final prediction. Conversely, if $\hat{e}^j_{t+1}$ is predicted to be small (also with respect to the loss of other base learners), $M^j$ will be important for the upcoming prediction. Even though the learning models comprising $M$ are trained in a batch way, the models in $\overline{M}$ are updated after every test observation. Moreover, the predictions by $\overline{M}$ are produced for each test observation, rendering an online nature to our method. Formally, we measure the weights of each base-learner using the following equation:

$$w^j_{t+1} = \frac{erfc(\hat{e}^j_{t+1})}{\sum_{i \in \overline{M}} erfc(\hat{e}^j_{t+1})} \tag{2}$$

where $\hat{e}^j_{t+1}$ is the prediction made by $\overline{M}^j$ for the absolute loss that $M^j$ will incur in $y_{t+1}$. The function $erfc$ denotes the complementary Gaussian error function which is formalized as follows:

$$erfc(x) = \frac{4}{\sqrt{\pi}} \int_x^{\infty} e^{-t^2} dt \tag{3}$$

The final prediction is a weighted average of the predictions made by the base-learners $\hat{y}^j_{t+1}$ with respect to $w^j_{t+1}$ computed according to Equation 4:

$$\hat{y}_{t+1} = \sum_{j=1}^{m} \hat{y}^j_{t+1} \times w^j_{t+1} \tag{4}$$

The arbitrating mechanism only stats to function after a period of *warm-up* of $\alpha$ observations. The reason for this is to provide robustness to the estimations of the losses $e$.

```
Algorithm 1: Arbitrated Ensemble for Water Consumption Forecasting
Input:
– Water consumption time series $Y$;
– External predictors $Y^{ext}$;
– embedding dimension $K$.
/* Learning Step */
Embed $Y$ and attach $Y^{ext} \rightarrow Y^K_{tr}$
foreach $M^j \in M$ do
    train $M^j$ using $Y^K_{tr}$
end
————————————————————————————————
    /* Metalearning Step */
for all available test observations $Y_{ts}$ do
    Embed $Y_{ts}$ and attach $Y^{ext} \rightarrow Y^K_{ts}$
    $\overline{X} \leftarrow$ Extract Metafeatures from $Y^K_{ts}$
    foreach $\overline{M}^j \in \overline{M}$ do
        train $\overline{M}^j$ to model: $e^j = f(\overline{X})$
    end
end
————————————————————————————————
    /* Predicting new observations */
foreach upcoming observation $y_{t+1}$ do
    Embed $y_{t+1}$ along with the $K-1$ most recent observations and attach $y^{ext}_{t+1}$
    Compute predictions $\hat{e}^j_{t+1}$ from models $\overline{M}^j \in \overline{M}$
    Compute weights $w^j_{t+1}$
    Compute predictions $\hat{y}^j_{t+1}$ from models $M^j \in M$
    Compute final prediction $\hat{y}_{t+1} = \sum^m_{j=1} \hat{y}^j_{t+1} \cdot w^j_{t+1}$
    Add $y_{t+1}$ to $Y_{ts}$ and return to metalearning step
end
```

Until the number of test observations reaches $\alpha$, the individual predictions of the base-learners are simply averaged according to the arithmetic mean. The proposed methodology is summarised in Algorithm 1.

## 4 Case Study

This study was conducted using data from the water distribution grid of the metropolitan area of Oporto, Portugal, which comprises around 1.7 million inhabitants. The system is maintained by *ADDP* [1] (*Águas do Douro e Paiva*) collecting both water consumption and pressure levels at several strategical points throughout the network with a temporal granularity of half an hour. The time span of the collected data ranges from 2012-01-01 to 2016-10-06.

### 4.1 Pre-processing

For simplicity we focused our work on a particular delivery point in the main city of Porto. The dataset is aggregated by day, reaching a total of 1741 observations across the above-mentioned time-span. The units of the consumption levels are in cubic meters ($m^3$).

In order to augment the information about the consumption levels we also collected weather data from a repository [31]. The information includes daily summaries about the mean and max temperature (in Celsius) and precipitation level (in millimeters) in the city of Porto. Moreover, we also considered information about special days (sundays and local holidays) and the mean and standard deviation of the most recent $K$ values as predictors. These are the external features comprising $Y^{ext}$.

Pressure levels across the distribution grid presented an inadmissible number of missing values and for that reason the information about pressure was discarded. Regarding water consumption levels, 5 out of the 1741 days had missing entries in the raw dataset of half an hour granularity. These were handled using a weighted nearest neighbors imputation technique. The exploratory data analysis carried out is presented in Figures 1 and 2. Figure 1 illustrates the water consumption levels per day of the year and respective standard deviation. The plot shows the overall daily dynamics and structure of the time series. In Figure 2 is depicted the daily water consumption by mean temperature, grouped by days that had any rainfall and days that had not. The LOESS curves suggest that in rainy days there is a lower water consumption than in days with-

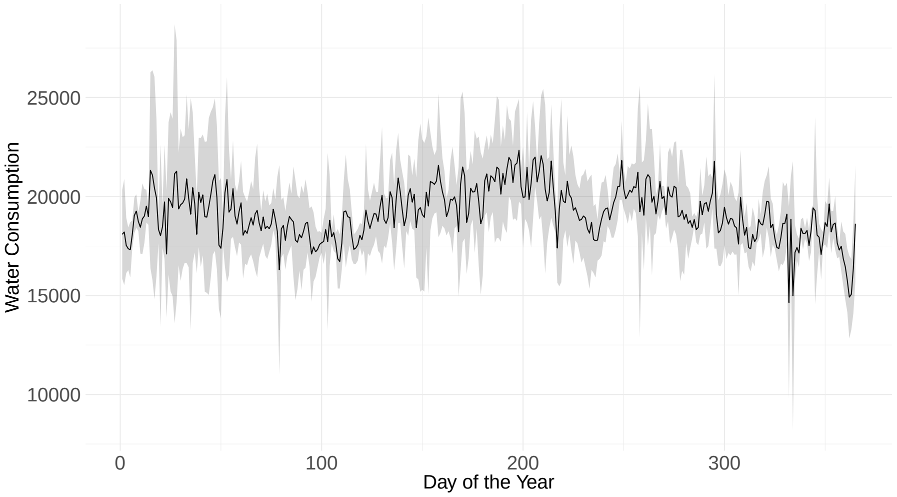


Fig. 1: Mean and standard deviation of water consumption ($m^3$) by day of the year

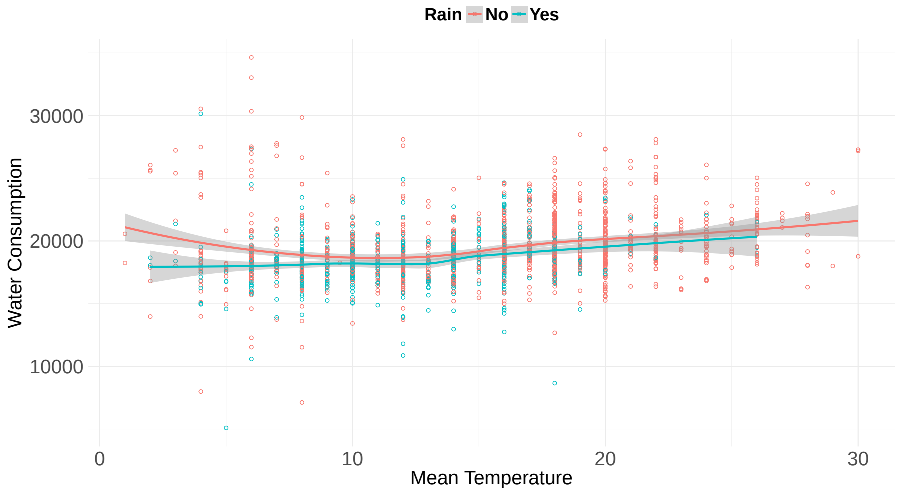


Fig. 2: Water consumption ($m^3$) by mean temperature ($C$), grouped by days that had or hadn't any rainfall. LOESS smoothing curves are included.

out rain. Moreover, it also suggests that the daily consumption increases as the mean temperature increases.

## 5 Empirical Experiments

This section describes the empirical experiments performed to validate the proposed method for STWC forecasting. These were designed to address to following research questions:

**Q1:** Is it beneficial to weight individual forecasters according to an Arbitrating scheme for STWC forecasting tasks?;

**Q2:** How does the performance of the proposed method relates to the performance of the state-of-the-art methods for STWC forecasting tasks?

The experiments were carried out using performanceEstimation [28] package. The methods used in the experiments were evaluated using the Root Mean Squared Error (RMSE) and the Mean Absolute Error (MAE) using a Monte Carlo procedure with **15** repetitions. For each repetition, a random point is picked from the time series. The previous window comprising **50**% of *t* is used for training and the following window of **20**% of *t* is used for testing.

The metafeatures used by $\overline{M}$ are the primitive ones, described previously in the Section 4.1, together with the following characteristics computed at each embedding vector: (**i**) **kurtosis**, which is a measure of flatness of the data distribution with respect to a

Gaussian distribution; (**ii**) **skewness**, which measures the symmetry of the distribution; (**iii**) **series trend**, calculated according to the ratio between the standard deviation of the series and the standard deviation of the differenced series; (**iv**) **serial correlation**, estimated using a Box-Pierce test statistic; and (**v**) **long-range dependence**, using a Hurst exponent estimation with a wavelet transform. These statistics summarise the overall structure of the time series of water consumption. For a comprehensive description of each statistic see [29].

We estimate the optimal embedding dimension ($K$) using the method of False Nearest Neighbors [12]. This method analyzes the behavior of the nearest neighbors as we increase $K$. According to the authors from [12], with a low sub-optimal $K$ many of the nearest neighbors will be false. Then, as we increase $K$ and approach an optimal embedding dimension those false neighbors disappear. $K$ is set to **6** in our experiments.

The $\alpha$ parameter for warm-up of $\overline{M}$ is **5**. Essentially, only after 5 observations does the arbitrating mechanism start to weight predictions. Until then, all base learners are averaged with equal weights.

The base-learners $M$ comprising the ensemble are the following: MARS [19], Generalized Linear Models [6], Random Forest [30], SVM [11], Rule-based regression [14], Generalized Boosted Regression [22] and Gaussian Processes [11]. Each of the individual learners is composed of 3 different parameter settings adding up to **21** learning models.

We tested two different algorithms as meta-learners: (i) a tree-based model and (ii) a Gaussian process with a radial basis function kernel.

We compare the proposed method to the following four baselines:

**ARIMA:** The state-of-the-art ARIMA model, using the function *auto.arima* from the forecast R package [8]. This function automatically tunes ARIMA to an **optimal parameter setting**;

**NN:** A feed forward neural network with a single hidden layer. The neural network was optimized using a grid search procedure using a total of 56 parameter combinations. The final parameter setting was 20 hidden units and a weight decay of 0.2;

**BT:** Bagged Trees from [20]. This bagging approach is specifically designed for time series forecasting tasks – in our experiments we used the reported best performing variant;

**S:** This is a variant of the proposed method, but stripped of the metalearning layer. That is, $M$ is trained in advance and their predictions are simply averaged at run-time using the arithmetic mean.

The results from the Monte Carlo experiments are reported in Table 1. Besides the baselines defined above, **E.DT** and **E.GP** denote the proposed **E**nsemble method, employing a decision tree and a Gaussian process as metalearners, respectively.

Relative to **S**, our approach is able to overcome its average performance, especially when using a Gaussian Process as metalearner. Nonetheless, the deviance of experiments are slightly higher in our proposed method for the RMSE. Overall, we conclude that indeed our metalearning approach is beneficial for our STWC forecasting tasks (**Q1**).

The performance of the proposed methods clearly overcome the performance of the state-of-the-art methods (**ARIMA** and **NN**) as well as other methods for time series forecasting tasks (**BT**) – which validates our hypothesis **Q2**.

In summary, our experiments validate our hypothesis that our proposed method is able to model the different dynamics of water consumption with a competitive performance relative to state-of-the-art methods.

| **Method** | **RMSE** | **MAE** |
|---|---|---|
| **ARIMA** | 1653.55±63.72 | 1199.82±61.80 |
| **NN** | 1805.96±273.60 | 1304.49±210.43 |
| **BT** | 1684.21±50.90 | 1220.64±41.24 |
| **S** | 1629.76±58.14 | 1179.72±60.61 |
| **E.DT** | 1622.47±63.50 | 1165.52±60.57 |
| **E.GP** | **1608.98±67.83** | **1162.45±60.19** |

Table 1: Average results from using RMSE and MAE with a Monte Carlo procedure

## 6 Conclusions

In this paper we presented a new method for STWC forecasting tasks. We argued that the planning of operations related STWC is an important topic with economical and social impact. Our proposed method settles on a metalearning scheme called Arbitrating, introduced before by [21] and [13]. We extend their ideas to STWC forecasting tasks.

We leverage the Arbitrating strategy to dynamically weight individual models in an ensemble. We reasoned that the series of water consumption follows a recurring pattern with different regimes. In effect, our approach allows a fast detection and adaptation to the different regimes causing the data.

Results from numerical experiments suggest that our metalearning is worthwhile. Moreover, we empirically demonstrate that the proposed method is competitive with other state-of-the-art techniques for STWC forecasting tasks, such as Neural Networks and the classical time series model ARIMA.

Future work includes: (**i**) generalize the proposed methodology for other time series forecasting tasks; (**ii**) Compare the proposed method against a time-dependent combining heuristic (e.g. recent performance of individual learners).

In the interest of reproducible research our methods are publicly available as an ***R*** package called **tsensembler**[1].

## References

1. ADDP: Oporto water consumption. `http://addp.pt`, accessed: 2016-11-21

[1] `https://github.com/vcerqueira/tsensembler`

2. An, A., Shan, N., Chan, C., Cercone, N., Ziarko, W.: Discovering rules for water demand prediction: an enhanced rough-set approach. Engineering Applications of Artificial Intelligence 9(6), 645–653 (1996)
3. Bougadis, J., Adamowski, K., Diduch, R.: Short-term municipal water demand forecasting. Hydrological Processes 19(1), 137–148 (2005)
4. Brazdil, P., Carrier, C.G., Soares, C., Vilalta, R.: Metalearning: Applications to data mining. Springer Science & Business Media (2008)
5. Brown, G.: Encyclopedia of Machine Learning, chap. Ensemble Learning, pp. 312–320. Springer US, Boston, MA (2010)
6. Friedman, J., Hastie, T., Tibshirani, R.: Regularization paths for generalized linear models via coordinate descent. Journal of Statistical Software 33(1), 1–22 (2010)
7. Herrera, M., Torgo, L., Izquierdo, J., Pérez-García, R.: Predictive models for forecasting hourly urban water demand. Journal of hydrology 387(1), 141–150 (2010)
8. Hyndman, R.J., with contributions from George Athanasopoulos, Razbash, S., Schmidt, D., Zhou, Z., Khan, Y., Bergmeir, C., Wang, E.: forecast: Forecasting functions for time series and linear models (2014), R package version 5.6
9. Jain, A., Varshney, A.K., Joshi, U.C.: Short-term water demand forecast modelling at iit kanpur using artificial neural networks. Water Resources Management 15(5), 299–321 (2001)
10. Jentgen, L.A., Kidder, H., Hill, R., Conrad, S., Papalexopaulos, A.: Water consumption forecasting to improve energy efficiency of pumping operations. AwwaRF, Denver (2007)
11. Karatzoglou, A., Smola, A., Hornik, K., Zeileis, A.: kernlab – an S4 package for kernel methods in R. Journal of Statistical Software 11(9), 1–20 (2004)
12. Kennel, M.B., Brown, R., Abarbanel, H.D.: Determining embedding dimension for phase-space reconstruction using a geometrical construction. Physical review A 45(6), 3403 (1992)
13. Koppel, M., Engelson, S.P.: Integrating multiple classifiers by finding their areas of expertise. In: AAAI-96 Workshop On Integrating Multiple Learning Models. pp. 53–58. Citeseer (1996)
14. Kuhn, M., Weston, S., Keefer, C., code for Cubist by Ross Quinlan, N.C.C.: Cubist: Rule- and Instance-Based Regression Modeling (2014), R package version 0.0.18
15. Lemke, C., Gabrys, B.: Meta-learning for time series forecasting and forecast combination. Neurocomputing 73(10), 2006–2016 (2010)
16. Maidment, D.R., Miaou, S.P.: Daily water use in nine cities. Water Resources Research 22(6), 845–851 (1986)
17. Maidment, D.R., Miaou, S.P., Crawford, M.M.: Transfer function models of daily urban water use. Water Resources Research 21(4), 425–432 (1985)
18. Mendes-Moreira, J., Soares, C., Jorge, A.M., Sousa, J.F.D.: Ensemble approaches for regression: A survey. ACM Computing Surveys (CSUR) 45(1), 10 (2012)
19. Milborrow, S.: earth: Multivariate Adaptive Regression Spline Models. Derived from mda:mars by Trevor Hastie and Rob Tibshirani. (2012)
20. Oliveira, M., Torgo, L.: Ensembles for time series forecasting. In: ACML Proceedings of Asian Conference on Machine Learning. JMLR: Workshop and Conference Proceedings (2014)
21. Ortega, J., Koppel, M., Argamon, S.: Arbitrating among competing classifiers using learned referees. Knowledge and Information Systems 3(4), 470–490 (2001), http://dx.doi.org/10.1007/PL00011679
22. Ridgeway, G.: gbm: Generalized Boosted Regression Models (2015), R package version 2.1.1
23. Rossi, A.L.D., de Leon Ferreira, A.C.P., Soares, C., De Souza, B.F., et al.: Metastream: A meta-learning based method for periodic algorithm selection in time-changing data. Neurocomputing 127, 52–64 (2014)

24. dos Santos, P.M., Ludermir, T.B., Prudencio, R.B.C.: Selection of time series forecasting models based on performance information. In: Hybrid Intelligent Systems, 2004. HIS'04. Fourth International Conference on. pp. 366–371. IEEE (2004)
25. Smith, J.A.: A model of daily municipal water use for short-term forecasting. Water Resources Research 24(2), 201–206 (1988)
26. Soares, C., Brazdil, P.B.: Zoomed ranking: Selection of classification algorithms based on relevant performance information. In: European Conference on Principles of Data Mining and Knowledge Discovery. pp. 126–135. Springer (2000)
27. Takens, F.: Dynamical Systems and Turbulence, Warwick 1980: Proceedings of a Symposium Held at the University of Warwick 1979/80, chap. Detecting strange attractors in turbulence, pp. 366–381. Springer Berlin Heidelberg, Berlin, Heidelberg (1981)
28. Torgo, L.: An Infra-Structure for Performance Estimation and Experimental Comparison of Predictive Models (2013), R package version 0.1.1
29. Wang, X., Smith-Miles, K., Hyndman, R.: Rule induction for forecasting method selection: Meta-learning the characteristics of univariate time series. Neurocomputing 72(10), 2581–2594 (2009)
30. Wright, M.N.: ranger: A Fast Implementation of Random Forests (2015), R package version 0.3.0
31. Wunderground: Weather forecast and reports. `https://www.wunderground.com/`, accessed: 2016-12-13
32. Zhou, S.L., McMahon, T.A., Walton, A., Lewis, J.: Forecasting daily urban water demand: a case study of melbourne. Journal of Hydrology 236(3), 153–164 (2000)
33. Zhou, S., McMahon, T., Walton, A., Lewis, J.: Forecasting operational demand for an urban water supply zone. Journal of Hydrology 259(1), 189–202 (2002)

# PAPERS IN ALPHABETICAL ORDER

# AUTHORS IN ALPHABETICAL ORDER

ISBN 978-972-752-207-1