Information Systems & Grid Technologies

Eleventh International Conference ISGT’2017

Sofia, Bulgaria, Sep. 29 – 30, 2017.

ST. KLIMENT OHRIDSKI UNIVERSITY PRESS

ISGT’2017 Conference Committees

Chair

Prof Vladimir DIMITROV

Program Committee

- Mícheál Mac an AIRCHINNIGH, Trinity College, University of Dublin
- Pavel AZALOV, Pennsylvania State University
- Irena BOJANOVA, University of Maryland University College
- Marco DE MARCO, Catholic University of Milan
- Milena DOBREVA, University of Malta
- Vladimir GETOV, University of Westminster
- Seifedine KADRY, American University of the Middle East, Kuwait
- Kalinka KALOYANOVA, University of Sofia “St Cl Ochridsky”
- Angelika KOKKINAKI, University of Nicosia
- Violeta MANEVSKA, University of Bitola “St Cl Ochridsky”
- Maria NISHEVA, University of Sofia “St Cl Ochridsky”
- Dov TE’ENI, Tel-Aviv University
- Stanislaw WRYCZA, University of Gdansk
- Fani ZLATAROVA, Elizabethtown College

Organizing Committee

- Vasil GEORGIEV
- Maria KOLEVA

Vladimir Dimitrov, Vasil Georgiev (Editors)

Information Systems & Grid Technologies

Eleventh International Conference ISGT'2017

Sofia, Bulgaria, Sep. 29 – 30, 2017.

Proceedings

organized by

Faculty on Mathematics and Informatics
University of Sofia St. Kliment Ohridski

Bulgarian Chapter of the
Association for Information Systems (BulAIS)

St. Kliment Ohridski University Press

Preface

This conference was being held for the eleventh time in the end of September, 2017 in the Faculty of Mathematics and Informatics in Sofia, Bulgaria. It is supported by the Science Fund of the University of Sofia “St. Kliment Ohridski” and by the Bulgarian Chapter of the Association for Information Systems (BulAIS).

Total number of papers submitted for participation in ISGT’2017 was 23. They undergo the due selection by at least two members of the Program Committee. This book comprises 15 papers of 13 Bulgarian and 13 foreign authors. The conference papers are available also on the ISGT web page http://isgt.fmi.uni-sofia.bg/ (under «Previous conferences» tab).

Responsibility for the accuracy of all statements in each peer-reviewed paper rests solely with the author(s). Permission is granted to photocopy or refer to any part of this book for personal or academic use providing credit is given to the conference and to the authors.

The editors

© 2017 Vladimir Dimitrov, Vasil Georgiev (Eds.)

ISSN 1314-4855
St. Kliment Ohridski University Press

Table of Contents

Aligning Behavior Competencies to the IS curricula
Kalinka Kaloyanova, Vijay Kanabar 7

Designing and Developing an Online Platform for 'Black Swan'
Events Management in Hospitality
Angelika Kokkinaki, Stella Kleanthous, Fotini Zioga, Kyrillou, Chrisostomi 17

A Brief Outlining of Entrepreneurship Restructuring as
Industrial Waves Occurred
Ioannis Patias 31

Challenges in Data Anonymity, Protection and Privacy in
Medical Research and Services
Anatoliy Dimitrov, Dimitar Vassilev 38

Integrated Semantic Data Model for Functional Annotation of
Protein Sequences
Deyan Peychev, Dimitar Vassilev 45

Discovery of Pattern Deviations Caused by Gene Duplication in
Plant Evolutionary Trees
Irena Avdjieva, Milko Krachunov, Ognyan Kulev, Dimitar Vassilev 54

Model of Information System for Company Management
According to Software Engineering Standards
ISO/IEC 12207:2013
Katerina Dimeska, Snezana Savoska 60

Big Data Visualization Concepts and Visual Data Analytics
Snezana Savoska, Dragan Milevski 69

BI Tools Analysis According to Business Criteria as
Data Integration Possibilities, Hardware Specification,
Tools for Data Visualization and Comparison of Used Technologies
Andrijana Bocevska, Snezana Savoska, Ivan Milevski....................80

Information Anonymization – Procedures, Politics and Techniques
Vladimir Dimitrov 91

Object Oriented Analysis and Design of Protein Fingerprints
Monica Dimitrova, Dimitar Vassilev 109

Pattern Classification of Alternative Splicing in Gene Expression
Ognyan Kulev 119

The Methodology Behind Applying Adapted QSAR as a Cost-Effective Method for In Silico Biotechnology Experiments
Valeriya Simeonova, Dimitar Vassilev... 125

Problem Space of the Monitoring Granularity in Cloud Systems
Vasil Georgiev, Daniel Simeonov, Hristo Hristov 136

Optimizing and Verifying Model Effectiveness Using Random Forest and Logistic Regression
Vladimir Nasteski, Violeta Manevska, Snezana Savoska 149

Author Index 159

Aligning Behavior Competencies to the IS curricula

Kalinka Kaloyanova[1, 2], Vijay Kanabar[3],

[1]Faculty of Mathematics and Informatics, Sofia University, 5 James Bourchier blvd.,
1164, Sofia, Bulgaria
[2]Institute of Mathematics and Informatics, Bulgarian Academy of Sciences,
Acad. Georgi Bonchev Str., Block 8, 1113, Sofia, Bulgaria
[3] Computer Science Department, Metropolitan College, 808 Commonwealth Ave.,
Room 250, Boston University, Boston, MA 02215 USA
kkaloyanova@fmi.uni-sofia.bg, kanabar@bu.edu

Abstract. *During the last two decades the researchers and practitioners in Information Systems (IS) area are constantly challenged by the changes in IT and the increasing role of information systems for the organizations. Also the education in this area is changed to reflect these challenges. Several curricula for graduate and undergraduate IS programs have been developed recently: IS 2002 and IS 2010 for undergraduate level and MSIS 2006 and 2016 for master students. Every curriculum tries to reflect the most challenging aspects of time. But also every curriculum requires a strong fundament in the area. Along with technical skills and knowledge in the field behavioral skills are no less important. In this paper we explore how behavior competencies are reflected in the curricula. We also present some other visions on the topic - the PMI view. Based on this study some conclusions are done and some recommendations are proposed.*

Keywords: Information Systems, IS curriculum, MSIS curriculum, competencies, behavior competencies

1. Introduction

During the last two decades the researchers, educators and practitioners in IS area have been seriously challenged by the chang es taking place in IT on several fronts. First across organizations the role of information technology and information systems in organizations continues to be very prominent. Additionally, due to the emergence of various digital innovations, such as apps on mobile devices, that provide services, and Internet of Things (IoT) technology, organizations are under constant pressure to change and innovate. This is where a new role, the *Innovative IS Project Manager* has started to emerge. For example, in the banking sector, the IS project manager might be asked to manage an initiative to reduce dependency of human bank tellers that support banking transactions such as deposit checks or transfer funds. The IS manager has to strategically describe how an IT project

V. Dimitrov, V. Georgiev (Editors): ISGT'2017. ISSN 1314-4855
Proceedings of the 11th International Conference on
Information Systems and Grid Technologies, Sofia, 29.-30. Sep. 2017.

involving mobile applications can improve customer satisfaction by 10% as well as deliver a quality product that does away with human operators. According to Gallagher [12], in this model, the project manager is part tactician—responsible for executing the scope of work in the time frame given—and part strategist, responsible for interpreting the business strategy, assessing feasibility of the objective, recommending and/or innovating a solution, formulating a scope of work, and executing the IT project. Additionally, in this model, the IT project manager plays a critical role in identifying a series of projects that provide benefits and then managing them. Education in both IS and Project Management (PM) has to change and reflect the new reality. As such several curricula for graduate and undergraduate IS programs have been developed recently, our paper will describe such changes so that educators can benefit from them.

The role of the prominent Association for Computing Machinery (ACM) needs to be acknowledged here The first ACM recommendations for IS curricula appeared in 1972, this was a modest start for IS curriculum. Several curricula in the recent years in particular – IS 2002 and IS 2010 for undergraduate level and MSIS 2006 and MSIS 2016 for master students are now filling a critical niche.

The IS curricula are result of the collaboration efforts of ACM and AIS (Association for Information Systems). The latest guidelines MSIS 2016 Global Competency Model for Graduate Degree Programs in Information Systems is truly exciting. Unlike its predecessors, which focussed on courses and modules, MSIS 2016 allows one to map curriculum with competencies.

In this paper we restrict our research and analysis to explore how behavioral competencies are reflected in the IS curricula. We also present another view on the topic - *Project Management Curriculum Guidelines* sponsored by the Project Management Institute (PMI). Finally, we give some recommendations from our analysis of these curricula and other sources [7, 12, 13].

2. Curricula and Guidelines in IS area

Since the early 2000 curricula in the IS area is jointly sponsored by ACM and AIS. This curricula represents the voluntary effort of numerous individuals, including faculty, practitioners and other stakeholders. This effort is grounded in the expected requirements for competencies from the industry and integrates the views of various organizations employing IS graduates [9]. The revisions of the IS curricula are usually spurred by different motivating factors: the Internet, IS accreditation movement, globalization, etc. Nowadays cloud computing, Big data and agile development provide opportunities but also pose new challenges for IS graduates.

2.1 IS 2002 Curriculum

The IS2002 curriculum was developed around several core dimensions of the IS professionals skills [2]:

- understanding the role of IS - a broad business and real world perspective
- technical skills – IS development, deployment and management
- strong analytical and critical thinking skills
- interpersonal skills (communication and team working skills).

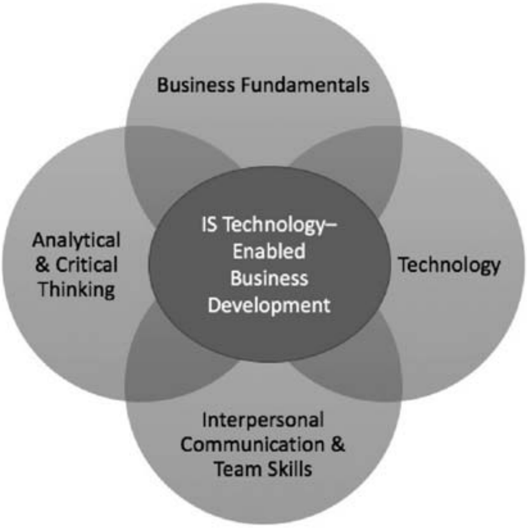

Business Fundamentals
Analytical & Critical Thinking
IS Technology–Enabled Business Development
Technology
Interpersonal Communication & Team Skills

Fig.1. IS 2002 Graduate Exit Characteristics Categorization
Source: [2]

In Figure 1 we illustrate the key domains such as Analytical & Critical Thinking, and Technology competencies. In this curriculum for the first time Project Management competencies are recommended This came about after a review of more than 63 programs – and many of them highlighted the importance of IT or Software Project competencies for IS graduates. Key issues underscored in the research were ability to lead software projects, work in teams, manage conflict, communicate & report results in a timely manner to stakeholders. This guideline specified 10 unique courses which one could adopt as is for teaching purposes.

2.2 MSIS 2006 Curriculum

The MSIS 2006 Model Curriculum and Guidelines for Graduate Degree Programs in Information Systems includes detailed descriptions for a set courses. For each course a name and course descriptions were presented [3].

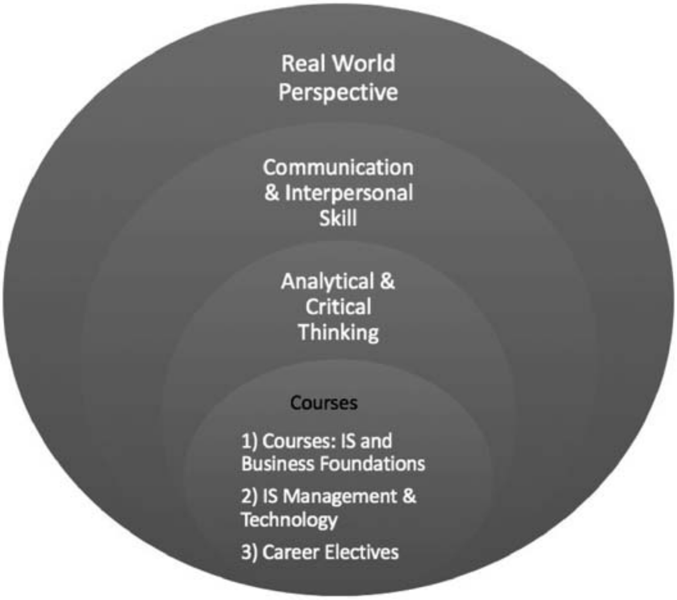

Real World Perspective
Communication & Interpersonal Skill
Analytical & Critical Thinking
Courses
1) Courses: IS and Business Foundations
2) IS Management & Technology
3) Career Electives

Fig. 2. MSIS 2006 –Skills, Knowledge, and Values
Source: [3]

The goals of the MSIS 2006 programs was described in the curriculum guidelines as follows [3]:

Students graduating from the MSIS program should be prepared to provide leadership in the Information Systems field and will possess the following skills, knowledge, and values upon graduation:

- A core of IS management and technology knowledge
- Integration of IS and business foundations
- Broad business and real world perspective
- Communication, interpersonal, and team skills
- Analytical and critical thinking skills
- Specific skills leading to a career.

Several career paths are suggested such as software development, data administration, systems integration, data communication, managing sourcing and global projects and project management. The guidelines communicated that with some customization, individual schools and departments can specialize their program offering in areas recommended above.

2.3 IS 2010 Curriculum

IS 2010 Curriculum represented the third collaborative effort by ACM and AIS [8]. It revised the earlier curricula, by providing more details on the learning outcome expectations for IS graduates [8,9].

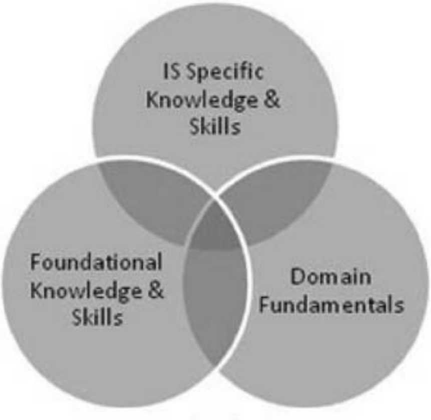

IS Specific Knowledge & Skills
Foundational Knowledge & Skills
Domain Fundamentals

Fig. 3. IS 2010 - High Level Curriculum Topics
Source: [9]

In addition, a set of seven courses was proposed as a core for all Information Systems Programs. These courses identify the specific content that defines information systems. See Figure 4 for the list of courses and their dependencies/ pre-requisities.

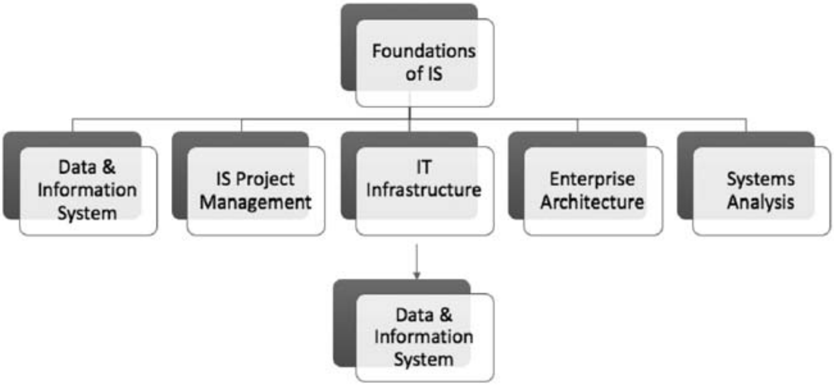


Fig. 4. IS 2010 Core Topics

Also a list of elective courses was included to extend the coverage provided by basic ones [1]. This guideline enhanced the careertracks based on groupings of electives. For example, if an institution was keen to offer a MS in IT Project Management, they would modify the curriculum as needed by adding two additional specialization courses in software PM. In this manner an institution can develop programs based on local market interest and faculty strengths.

2.4 MSIS 2016 Curriculum

The last updates in IS area are reflected in the MSIS 2016 Global Competency Model for Graduate Degree Programs. Unlike the previous ones, this curriculum focuses on IS competencies rather than just prescribing a list of courses and knowledge areas [10,11].

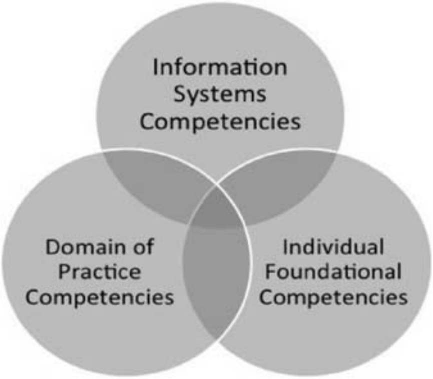

Information Systems Competencies
Domain of Practice Competencies
Individual Foundational Competencies

Fig.5. MSIS 2016 - High Level Realms
Source: [11]

According to this curriculum, an MSIS program should develop competencies within the next several subjects [10]:

- Computing Competencies
- IS Management Competencies
- Individual Foundational Competencies
- Domain Competencies.

The nine MSIS 2016 core Computing/IS Management competency areas are: Business Continuity and Information Assurance; Data, Information, and Content Management; Enterprise Architecture; Ethics, Impacts, and Sustainability; Innovation, Organizational Change, and Entrepreneurship; IS Management and Operations; IS Strategy and Governance; IT Infrastructure; and Systems Development and Deployment [11].

With focus on Innovation, Organizational Change, and Entrepreneurship the curriculum is fresh enough to prepare students for the digital economy. Students can learn new business models such as what we have seen from the likes of UBER and AirBnB and their impact on organizations.

MSIS 2016 also recognizes that there are other disciplines and industry domains that might benefit from MSIS curriculum. The earlier curricula assumed that the primary clientele is Business Schools and Information Systems or Computer Information Systems (CIS) departments. 2016 explicitly recognizes that business is not the only area of interest for the IS domain. For example, health informatics can now be accommodated in the new curriculum guideline.

3. Behavior Competencies in Curriulum Frameworks

In MSIS 2016, the model specifies a set of prerequisite competencies. Some of these competencies, like technical competencies, are included into the competency areas and specified for MSIS graduates. Additionally it is expected that a lot of technical competencies will be inherited from the previous, undergraduate level course work by students.

Let us address the soft skills competencies supported in the MSIS 2016. For instance, collaboration and team work are important.

With regards to individual foundational competencies, there are no specific recommendations. It is assumed that the students will have some level of ability in communication (oral and written), as well in leadership, collaboration, negotiation, etc. [11]. However, the guidelines references and recommnds the following attributes based on Stevens and Campion [13] for teamwork:

1) conflict resolution
2) collaborative problem solving
3) communication
4) goal setting and performance management
5) planning and task coordination.

To note, the MSIS 2016 Global Competency Model for Graduate Degree Programs in Information Systems includes individual foundational competencies as a major area in the IS education, but it does not explain how students will acquire them. Instead, it references European e-Competence Framework, where a fluid list of 88 examples of competency is discussed.

Nevertheless, the IS guidelines are not sufficiently detailed about the competency areas in project management (PM), some examples of competency that pertain to PM areas could be used [5]: Managing IS projects and programs, Managing IS project portfolios and Managing IS development projects.

For these three areas the Knowledge Modules (KMs) in the behavioral domains may be implemented into the IS curricula:

- Plan, Distribute, and Manage Project Communications
- Project Team Building and Motivating
- Project Leadership

- Identifying and Engaging Stakeholders
- Project Organization and Context
- Managing Global Projects
- Virtual Project Management
- Ethics and Professionalism

Each KM further could be detailed into ten knowledge topics and three learning outcomes. Instructors can embed them into the courses as needed, using different approaches [4,6].

If faculty are keen to provide comprehensive strengths in these domains we strongly recommend that the curriculum designers download a valuable exemplar course available for free to academic at http://www.pmiteachorg. The details of PM competencies can be found in the course PM-2: Project Teams, Leadership, and Communications (Task Force on PM Curricula (2017) [14]). The students studying the recommended topics will be able to:

1. Describe how project purpose, organizational structure, culture, roles, and knowledge management impact their ability to lead a project team.

2. Identify, recognize, and engage both internal and external project stakeholders.

3. Explain roles of team members and how they contribute to projects, and apply an understanding of self in how they relate to teams.

4. Apply knowledge of team building, operation, and behavior, considering factors that influence effectiveness for all types of teams, including those that are virtual and global.

5. Identify and apply an appropriate communication strategy for a given situation.

6. Analyze power dynamics and organizational politics, and suggest conflict management and negotiation techniques relevant to the project.

7. Apply critical and innovative thinking skills to improve their own, and others’, decision-making processes within projects.

8. Describe and relate basic leadership principles and processes to address leadership issues and to remove obstacles in the project environment.

9. Describe the organizational change management process and encourage behavior to maximize project success.

10. Describe underlying ethical principles, which should be applied to managing projects, and demonstrate ethical sensitivity for a given situation.

4. Challenges and Opportunities

In the light of the above enhancements to information systems and project management curricula, let us examine some challenges and opportunities. First, there is a little opportunity to continue to introduce new dedicated courses dealing with behavioral competencies in most computer science and information systems curricula as noted in [6]. Given an opportunity to introduce new courses, IS faculty would prefer to integrate courses such as cybersecurity or data analytics into their program. So even though we wish a course like PM-2 becomes mandatory in IS curriculum, we are not optimistic it will happen in most programs.

However, there is a trend towards modularization in curriculum design these days. For instance, at MIT (see http://news.mit.edu/2014/future-of-mit-education-0804), the task force championed by the president lays down the following vision:

"The MIT education of the future is likely to be more global in its orientation and engagement, more modular and flexible in its offerings, and more open to experiments with new modes of learning …… the report suggests that students are focused more on learning certain elements of a class and less on completing what has traditionally been considered a module or unit of learning."

Such modularization might provide unique opportunities from two perspectives. First, it is possible that larger courses and entire programs will be chunked in a flexible manner. If so, there will certainly be room to introduce units such as *Plan, Distribute, and Manage Project Communications* and *Project Leadership* into an information systems program. Second, entire programs might be compressed and a fully dedicated course dealing with behavioral competencies might be introduced.

Finally, let us the consider the opportunity, that all most all information systems curricula have one or more courses that support term papers, capstone projects which are project driven. In most such cases students have to work in teams towards some real-world tangible deliverable. We strongly recommend that all such courses should embed the fundamentals of behavioral competencies. It is desirable to have our students successfully completing external projects, and keeping the external stakeholders who provide IS projects for the teams satisfied. One way to assure that is to make sure that the students are competent at communications, leadership and stakeholder engagement.

5. Conclusion

The Information Systems curricula have constantly evolved over the years in a positive maner. The different revisions reflect the challenges of the changing technologies and the needs of the industry and how faculty globally are reacting by responding to changes. The MSIS 2016 is a very flexible competency based guideline, which is very valuable for many institutions in many different academic units ranging from business and computer science to health sciences.

In this paper we have highlighted a key concern about focusing only on technological trends and skills of the IS graduates. The curriculum needs to address how we can develop behavioural competencies of students. IS students will be working in teams, and leading transformative projects. We addressed how some of these aspects were reflected in the latest IS curricula. We also gave a short analysis how the PM-2 curriculum guideline can be invaluable and we recommend using elements from this standard as well. Our next research focus is to describe how the MIT framework approach or PM-2 approach can be integrated with MIS 2016 to develop a useful curriculum.

Acknowledgment. This paper is supported by Sofia University "St. Kliment Ohridski" SRF under Contract 80-10-119/2017.

References

1. Bell C., R. Mills, and K. Fadel, "An Analysis of Undergraduate Information Systems Curricula: Adoption of the IS 2010 Curriculum Guidelines," Communications of the Association for Information Systems, vol. 32, no. 2, (2013)
2. Gorgone J. T, et al, IS 2002, Model Curriculum and Guidelines for Undergraduate Degree Programs IS: Association for Information Systems, (2002)
3. Gorgone J.T., P. Gray, E. A. Stohr, J.S. Valacich, and R. T. Wigand, MSIS 2006. Model Curriculum and Guidelines for Graduate Degree Programs in Information Systems, Communications of the AIS, Volume 17, No 1 (2006)
4. Kaloyanova K., An Implementation of the Project Approach in Teaching Information Systems Courses, In Proceedings of the 8th International Technology, Education and Development Conference, Valencia, Spain, pp. 7090-7096 (2014)
5. Kanabar, V. and C. Messikomer: Design and Implementation of an Adaptive Curriculum Framework for Project Management Education, In Proceedings of IRNOP 2015, International Research Network on Organizing by Projects, London, UK (2015)
6. Kanabar, V., K. Kaloyanova, Identifying and Embeding behaviour competencies in IS courses, ECIS In Proceedings of the 25th European Conference on Information Systems (ECIS), Guimarães, Portugal, June 5-10, 2017 (pp. 3115-3122). ISBN 978-0-9915567-0-0, ECIS 2017 Proceedings, http://aisel.aisnet.org/ecis2017_rip/59/(2017)
7. Maneva, N., N. Nikolova. Soft Skills Training for Software People, Proc. of the 7-th Int. conf. CSECS, July 6-10, Dobrinishte, Bulgaria, 2011, ISSN 1313-8624, pp. 117-129 (2011)
8. Topi H, J. et al, IS 2010 Curriculum Guidelines for Undergraduate Degree Programs in Information Systems (2011)
   http://www.acm.org/education/curricula/IS%202010%20ACM%20final.pdf
9. Topi H. et al., "IS 2010: Curriculum Guidelines for Undergraduate Degree Programs in IS," Communications of Information Systems, vol. 26, (2010)
10. Topi, H., Brown, S. A., Carvalho, J., Donnellan, B., Karsten, H., Shen, J., Tan, B. C. Y. & Thouin, M. F. MSIS 2016: a comprehensive update of graduate level curriculum recommendation in Information Systems, AMCIS 2016 Proceedings of the 22nd Americas Conference on Information Systems (AMCIS) (pp. 1-3) (2016)
11. Topi H, H. Karsten, S. Wrown, J. A. Carvalho, B. Donnellan, J. Shen, T. bernard, M. Thouin, MSIS 2016: Global Competency Model for Graduate e Degree Programs in Information Systems, (2017)
    https://www.acm.org/binaries/content/assets/education/msis2016.pdf (2017)
12. Gallagher, S. Time, risk, and innovation: creating space in your day to solve meaningful problems. Paper presented at PMI® Global Congress 2015—EMEA, London, England. Newtown Square, PA: Project Management Institute, (2015)
13. Stevens, M.J., & Campion, M.A. The knowledge, skill, and ability requirements for teamwork: Implications for human resource management. Journal of Management, 20, 503-530, (1994)
14. Task Force on PM Curricula, PM curriculum and resources (Vol 3). Newtown Square, PA: Project Management Institute, (2017)

Designing and Developing an Online Platform for 'Black Swan' Events Management in Hospitality

Kokkinaki, Angelika, University of Nicosia, Nicosia, Cyprus,
kokkinaki.a@unic.ac.cy

Kleanthous, Stella, University of Nicosia, Nicosia, Cyprus,
kleanthous.s@unic.ac.cy

Zioga, Fotini, Open University Cyprus, Nicosia, Cyprus,
zioga.f/@ouc.ac.cy

Kyrillou, Chrisostomi Maria, University of Nicosia, Nicosia, Cyprus,
cmkyrillou@gmail.com

Abstract Black Swan Events or Low Probability High Impact Events (LoPHIEs), like wildfires, earthquakes and volcanic eruptions have significant implications to hospitality and travel industry; a vivid demonstration of the complexity of interconnections between organizational units involved in or affected by LoPHIEs is outlined in this paper through a case study. Appropriate preparations for hospitality and travel SMES to handle such events are identified and implemented through a purpose specific information system. To design this platform we followed a requirements analysis methodology. The main elements from the implementation of this platform have also been included whereas elements of an initial usability evaluation have also been presented.

Keywords: Black Swan Events, Hospitality, Platform Application.

1 Introduction and Research Background

Black Swan Events or Low Probability High Impact Events (LoPHIEs), which include crises, disasters or emergencies, have significant implications upon the operation of involved and affected organizations (Hergert, 2004); the hospitality and tourism industry is directly affected by LoPHIEs. Wildfires, earthquakes and volcanic eruptions demonstrate the complexity of interconnections between organizations involved in or affected by LoPHIEs. Appropriate preparations to such events may make the difference between a major disruption of operations in the affected organizations or their resilience and survival (Coombs & Holladay 2010; Halder, 2015). Organizations' preparedness to LoPHIEs is usually distinguished into three phases, namely methods that prepare the organization BEFORE the event; methods that are initiated DURING the event to limit damage and methods that examine the aftermaths (Bernstein, 2011; Coombs 2007; Coombs & Holladay, 2010).

Such approaches exhibit some fundamental limitations (Diakou and Kokkinaki 2013). The most common limitation in these approaches is the biases raised in judgment or decision making that usually have a huge impact in the quantification of probability, uncertainty and risk (e.g. Armstrong, 2006; Berg et al., 2009; Fildes et al., 2009; Goodwin & Wright, 2010; Jakoubi & Tjoa, 2007; Onkal & Gonul, 2005; Pennock et al., 2001). This proved to be especially important when an event occurs and has an impact to a specific business sector for example the travel industry. Consequently, we need to understand the procedures and needs of this sector and how do they currently respond to possible crises like the eruption of Eyjafjallajökull in Iceland.

The paper is structured as follows: Section 2 will describe the requirements analysis methodology for designing the online platform for black swan events management, section 3 will give information and description of the online platform and section 4 will provide an initial usability evaluation of the platform. Section 5 will conclude this paper.

2 Methodological Procedure for Requirements Analysis

The purpose of this study was to understand how a company in the hospitality and tourism industry handles the effects of LoPHIEs and through this to identify the requirements for a web information system that can be used in such situations from small businesses in the tourism industry and specifically tourist accommodation.

We followed a qualitative approach that involved semi-structured interviews with different people working in the company to understand what they know about Black Swan events and what measures the company took when they had to deal with such an event in past.

This empirical research was performed at a company in Europe, specializing in tourist lodging and accommodation services. One of the basic services of the company is its customer service and support, available in over 40 languages, along with support in the hotel/accommodation owners who are using the platform in a daily basis for updating accommodation availability. The company is showing rapid growth in recent years in response to serving travellers and accommodation providing companies, as well as its continuous optimization on-line platform to better serve their users.

The choice of this company was based on its worldwide impact and its active presence on one of the biggest Black Swan Events. It also combines experienced staff in both technology and administration. All the above in combined with the easy access we had to this company led to its choice as an environment of our research.

2.1 Instruments and Procedures

To select the sample of employees from the aforementioned who would be involved in the survey, a sampling frame was created based on specific criteria. These predetermined criteria are intended to give us as accurate results as possible:

More specifically, the introduction of the above criteria includes:

Employment time: depending on how long they are working in the company and have faced different events that they could be considered as "Black Swan" events. So, they can report/outline these events and analyze what they themselves perceive as more important.

Field of Specialization: the specialization of each employee is of great importance for how perceives, events and what information he can provide on "Black Swan" events.

Department: the department to which the employee belongs gives us information about the collective reaction of the department to that event. Also, for the changes that have been applied to that department for addressing Black Swan phenomena in the future.

The sample of the research is targeted and very carefully selected to be able to provide answers to the questions of this research. This approach helped to collect information about what is considered a "Black Swan" event in different parts of the company and to triangulate the collected data within the company's various departments. According to the criteria established, the sample of the survey is presented to the table below.

Table 1 Interviews undertaken with employees

Code No.	Position/Department	Position/Department During the event	Employment time
E1	Senior Coordinator / Marketing	Global Support Coordinator / Hotels Department	6
E2	Global PR Coordinator / Personal relationships	Content Editor / Content Department	5
E3	Team Leader Long Tail Support / Strategic Partnerships	Project manager / Customer Support	11
E4	Account Manager XML / IT	Senior Customer Care / Customer Support	10
E5	Senior Team Leader/ IT	Back-end Developer / IT	10
E6	Product Owner / Marketing	Technical Coordinator / Strategic Partnerships	8

After we selected the sample for this research we followed a procedure mandatory for complying with the ethical considerations of the research and the procedure of the company. Firstly, we informed the human resource department that provided the primary investigator with the permission to perform the study within the company and using the specific sample. Then the selected employees have been invited via email to participate in the study, provided them with details about the purpose of the study, the procedures that would be followed, and the duration of the interviews, and asked their permission for using a recording device during the interview. Most of the participants agreed to the interview recordings and for those who did not agree only transcripts were used.

2.2 Case Study - Eyjafjallajökull Volcano Eruption

This section aims to present the findings of the primary data collection phase conducted at a well-known accommodation provider company regarding "Black Swan" events. More specifically, the section provides a description and analysis of the data collected through interviews with the employees of the company regarding their perception of which events are considered as "Black Swan" events, who are called to deal with these events upon their occurrence, how the corporation reacts to them and finally how its organizational structures can be affected.

The sample of this study has been selected in such a way as to effectively achieve disparity across different departments within the company and responsibilities of the employees. The objective was to collect as much information as possible about what is considered as a "Black Swan" event, from different sections of the company. Initially we needed to investigate the familiarity of the participants and employees in the company with Black Swan events and their experiences with such events.

Through the analysis of the data gathered, it can be concluded that only one employee in the company was familiar with the term ''Black Swan" when describing an event. In the course of the interviews however, it turned out that while there have been several events experienced by the employees that meet the characteristics of Black Swan events, the term is not used within the company to describe them. During the interviews, most of the respondents reported different events that occurred during the time of their employment in the company and had the characteristics of a black swan. After free and open communication between the researcher and the respondents, the eruption of the Icelandic volcano, *Eyjafjallajökull, which took place in March 2010*, was considered as a representative example of a black swan event that most of the employees were familiar with, consequently, it was selected for further analysis.

The eruption of the Eyjafjallajökull volcano led to the gradual closure of Europe's airspace for six days in order to avoid airplane crashes. European cities were therefore left with closed airports and all passengers traveling to destinations outside Europe, with intermediate stops at European airports, were also affected. Based on statistics, it is estimated that the event affected 105,000 flights in total and 7 million passengers in many different countries for almost one month (21 March to 24 April). In addition, the economic impact of the volcanic eruption was extremely high, causing loses amounting to $ 5 billion due to airspace closure, tourism revenue and productivity decline. Based on estimations, the explosion caused damages on insured property amounting to $98 million, ranking it as the sixth most devastating explosion.

2.2.1 Implications of the event

As expected, such an event affected the industry stakeholders (customers and hospitality providers), as well as intermediators between hotel tenants and hoteliers. According to respondents, the crisis that the company had to deal with is summarized in the following points:

- Large number of accommodation cancellations by passengers who could not reach their destination.
- Passengers who had to extend their accommodation stay due to the closure of airports.
- Company employees who could not return to their base.

As a result of the above, the number of requests for amendments and cancellation of reservations increased dramatically. Initially, the Customer Service Department was the division directly affected by the situation, due to the insufficient number of employees to manage effectively the significant increase in the call volume and online requests. However, as the LoPHIE unfolded it became clear that all organizational units have been affected as outlined in the sequel.

2.2.2 Addressing the event

Based on the Eyjafjallajökull volcano eruption, the following sub-sections describe a series of actions for addressing the event, as emerged through interviews with company employees and aim to form the foundations of a consolidated action plan for dealing with Black Swan events within the tourism industry.

2.2.3 Impact Group

In this chaotic situation, action was required to achieve stabilization and resolution of problems inefficient manner. As respondents reported, one of the first steps taken was the setup of an impact group for coordinating actions and regulating the situation both internally and externally. This group was attended by individuals from each department. From an organizational point of view, the impact group consisted of people who either had management posts, or acted as project managers, or were team leaders or technicians.

The way the impact team worked was critical for decision-making. Group meetings were scheduled every two hours (approximately) during the day. Initially, each unit represented in the impact group provided information on the evolution of the events, the problems they were facing and what they had been successfully resolved. Then, there was an open discussion and effort to solve problems using brainstorming.

It is noteworthy that the company maintained a powerful IT department that could provide solutions to address this crisis by building new components in IT systems. That is why the participation of computer technology in the impact group meetings was critical for several reasons. Indicatively, they could respond at the same time to what configurations are feasible, how long they thought it was needed for implementing those configurations, and to suggest a different approach if they felt it necessary.

Subsequently, the participants in the impact group communicated the decisions taken to the other employees of the departments by summoning brief meetings and sending emails. The procedure described above was repeated until a relative stability occurred.

2.2.4 First line of communication

One of the decisions taken by the impact group was the establishment of an emergency communication line; representatives handling calls through this line undertake the decision if the issue presented is trivial and proceed with its address or it presents complications. In the latter case, it is forwarded to another dedicated communication owed. All personnel were updated and asked if they wanted to join the emergency communication line. The second line of communication was handled by employees in the Customer Service department.

Until the instantiation of this specific LoPHIE, only customer service employees were trained on the procedures to be followed in emergency situations. In order to gain competencies addressing customers' requests effectively one ought to attend a thorough training session which lasted a month. Due to time shortage, the Personnel Training Department had to reconsider and parameterize the educational material focusing only on typical procedures and create short intensive training seminars for employees serving the front-end emergency line of communication.

For optimal implementation of the emergency line initiative, in-house chat rooms were created, where employees could ask for clarifications about the case they had to handle from more experienced personnel. This enabled the speedy exchange of formal and tacit knowledge between novice and experience customer representatives. In addition, call tracking systems were further developed in the company's call centre and record the volume of calls received by each group. This made it easier to locate overloaded support lines and call forwarding could be then assumed. During this event all employees involved in resolving and restoring stability had to work overtime.

2.2.5 Extending Customer Service

Despite the immediate mobilization of the company to help the customer service department, the volume of requests was too high and additional measures were needed to cope with the emergency. The company made mass recruitment for the Customer Service department. Within one day, 100 people were recruited, with a monthly contract and started to work the next day. To enable the newly recruited employees, handle their duties on the next day, the Personnel Training Department created a second round of high-speed training seminars. One of the objectives of the high-speed training seminars, was also to enable them to understand the philosophy of the company before taking part in the crisis management.

2.2.6 Negotiations with hotels

A special weight in the management of this crisis was the negotiations that took place with the hotels regarding cancellation of reservations made through the company's website without any cancellation costs. An excellent cooperation between hotels and the company has been reported, with no intention of exploiting the situation. In most cases, cancellations were made free of charge and where this was not possible; hoteliers offered the tenants the option of rescheduling their booking for another time in the same or another hotel of the same standards that they were collaborating.

2.2.7 Contacting hotel tenants

According to the respondents, the company was not active in social media network, at the time of the event; something that made it difficult to connect and communicate instructions to hotel occupants who had not yet contacted the Customer Services department. In cooperation with the IT department, an extra space was created on the homepage of the website, where instructions were posted to hotel tenants regarding what they should do if their reservations were affected by the volcano eruption.

Essentially, it was defined and advised that the arrival of the tenants was considered as "urgent" if it was within the next 45 hours, where priority was given to settling the cancellation or extension of their stay. In this way an initial routing of the communication was established, and the additional calls could be avoided.

2.3 Organizational structures of the company after the event

Responses determining the recovery time are found to be indistinguishable, differing based on the department and the position held by the respondents. The departments that took longer to return to normal work were the Customer Service and the Hotel Support departments, as these were the departments most affected by the crisis. The recovery period was reported to range from 2 to 5 weeks.

According to the same respondent, the additional components of the information systems created only to support this event had to be re-implemented to be included as core functions in the systems, which was quite time consuming.

The results in the table below are presented in more detail.

The participants to the study pointed out that this was an extreme, unpredictable event and that the company was not ready to deal with it. Nonetheless, it is encouraging that the company was organized; the departments involved worked coordination and reacted quickly to the situation, thus achieving to provide support to both tenants and hotel owners.

Table 2 Recovery timeout

Department	**Reset Time**
Marketing	2 weeks
Customer Service	4-5 weeks
Hotel Support	4 weeks
Content	2 weeks

When the situation was stabilized; meetings were held in order to evaluate the way the company reacted and what could have been done differently if such a situation happened again in the future. A common conclusion from these meetings was the need to establish procedures to be followed in emergencies. In addition, one further conclusion was that the information systems used within the company should be scalable and more flexible.

Part of the procedures established after the eruption of the volcano were concerned with the creation of the first communications line by the other departments in order to provide support to the Customer Service Department in cases of unusually increased number of calls (i.e. high season demand in the summer months). In achieving this, the way in which staff training is carried out has been restructured. Furthermore, the IT department had to set up the appropriate infrastructure to enable calls and call center support requests to be routed to the rest of the departments.

The Eyjafjallajökull crisis helped the company redefine how customer service is managed. Over the next years, more emphasis has been put on enhancing the company’s customer-centric philosophy. Customer Service department has been restructured; new roles have been identified to improve communication and new processes have been implemented to enable settlement of similar events more effectively in the future.

3 Online Platform for Black Swan Event Management

Based on the insights outlined in the previous section, it appears that certain functional specifications are considered as important for the organization to be able to handle a “Black Swan” event. This is linked both to their immediate reaction of employees as well as the direct parameterization of information systems that are used to make it easier to deal with the situation immediately. Our goal is to develop an online management application that will be used from small-sized hotel units, which apart from the basic functionality, will be enriched with additional “Black Swan” events management mechanisms.

For the purpose of demonstrating the proposed Event Management Services for “Black Swan” events, basic hotel management mechanisms were also implemented. Through the online application owners or administrators will be able to manage the reservations and the hotel’s availability as well as access the guests’ details for allowing further communication with them. The additional “Black Swan” events management mechanisms enable managers of the hotel unit to follow simple procedures that can face a potential crisis, without the need for additional technical support. Since any additional technical support for information systems in small businesses is not covered by their own employees, but instead by external partners, this online application aims to address issues with minimal response time to any LoPHIE.

The application consists of the following:

Back-end that contains the management part of the application e.g. room availability, and the management of additional procedures related to Black Swan events.

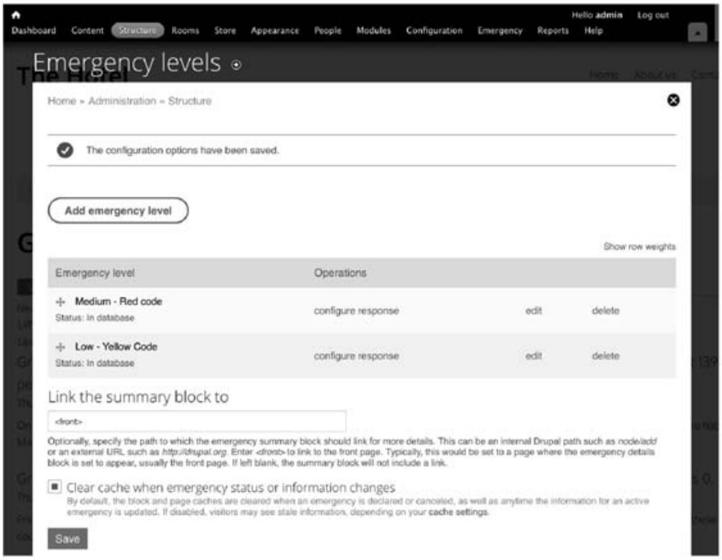
Front End that consists of the public space of the application that can be used by the customers for booking a room and get additional information on the hotel and the surrounding area.

3.1 Back-End Implementation based on Requirements

For the development of an online platform that could be employed to address LoPHIs, we used the Drupal CMS. In this paper we will focus only on the design and implementation of the Back-end features that contain also the tools for Black Swan events management.

3.1.1 Hiding the room search mechanism

For the implementation of the mechanism that will hide the search engine in case of emergencies from the public site; in lieu the Emergency module has been used. This module allows for adapting the public space of the website in case of emergencies (Figure 1). It gives the flexibility of creating multiple emergency levels according to the situation. For this application we created two levels (Medium and Low) and are both related to the search for availability.



Emergency levels

Home » Administration » Structure

The configuration options have been saved.

Emergency level	Operations		
Medium - Red code Status: In database	configure response	edit	delete
Low - Yellow Code Status: In database	configure response	edit	delete

Link the summary block to

<front>

Optionally, specify the path to which the emergency summary block should link for more details. This can be an internal Drupal path such as *node/add* or an external URL such as *http://drupal.org*. Enter <front> to link to the front page. Typically, this would be set to a page where the emergency details block is set to appear, usually the front page. If left blank, the summary block will not include a link.

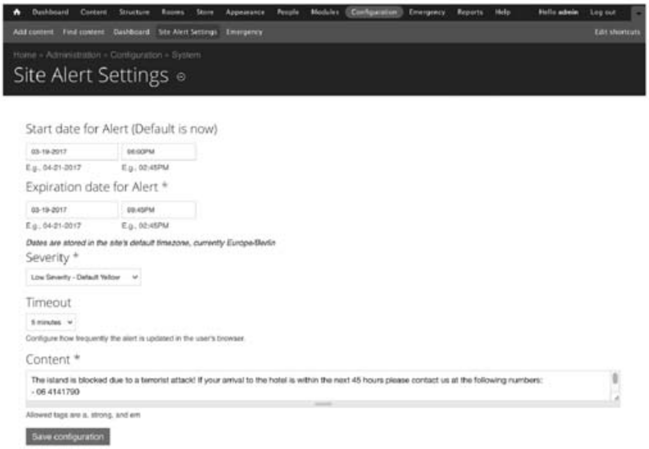
☑ Clear cache when emergency status or information changes

By default, the block and page caches are cleared when an emergency is declared or canceled, as well as anytime the information for an active emergency is updated. If disabled, visitors may see stale information, depending on your **cache settings**.

Figure 1 Emergency Module for hiding room search

3.1.2 Implementation of Alert messages

The site Alert (Figure 1) module has been used for implementing alert messages that would appear in case of an emergency. This module allows for setting the timeframe for the alert messages to appear and disappear after the time frame ends.


Dashboard Content Structure Rooms Store Appearance People Modules Configuration Emergency Reports Help
Hello admin Log out
Add content Find content Dashboard Site Alert Settings Emergency
Edit shortcuts
Home » Administration » Configuration » System
Save configuration

Site Alert Settings

Start date for Alert (Default is now)
03-19-2017 06:00PM
E.g., 04-21-2017 E.g., 02:45PM

Expiration date for Alert *
03-19-2017 09:45PM
E.g., 04-21-2017 E.g., 02:45PM

Dates are stored in the site's default timezone, currently Europe/Berlin

Severity *
Low Severity - Default Yellow

Timeout
5 minutes
Configure how frequently the alert is updated in the user's browser.

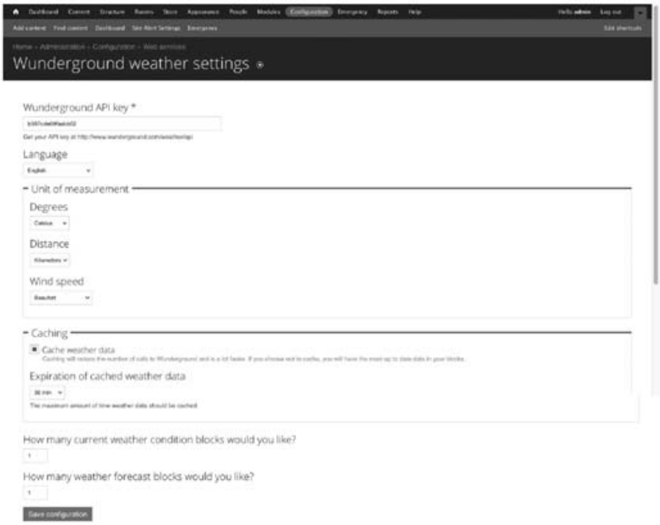
Content *
The island is blocked due to a terrorist attack! If your arrival to the hotel is within the next 45 hours please contact us at the following numbers:
- 06 4141790

Allowed tags are a, strong, and em

Figure 2 Alert Messages

3.1.3 Update messages from relevant parties

Updates from stakeholders are made using the Web Services Drupal (Wunderground weather module and Feed Aggregator see Figure 1), which are projected at the management panel through block structures. The manager of the hotel unit can customize the sources Updating by creating and adding new feeds.


Dashboard Content Structure Rooms Store Appearance People Modules Configuration Emergency Reports Help
Hello admin Log out
Add content Find content Dashboard Site Alert Settings Emergency
Edit shortcuts
Home » Administration » Configuration » Web services
Save configuration

Wunderground weather settings

Wunderground API key *
[illegible]
Get your API key at http://www.wunderground.com/weather/api

Language
English

Unit of measurement

Degrees
Celsius

Distance
Kilometers

Wind speed
Beaufort

Caching

■ Cache weather data
[illegible]

Expiration of cached weather data
30 min
The maximum amount of time weather data should be cached

How many current weather condition blocks would you like?
1

How many weather forecast blocks would you like?
1

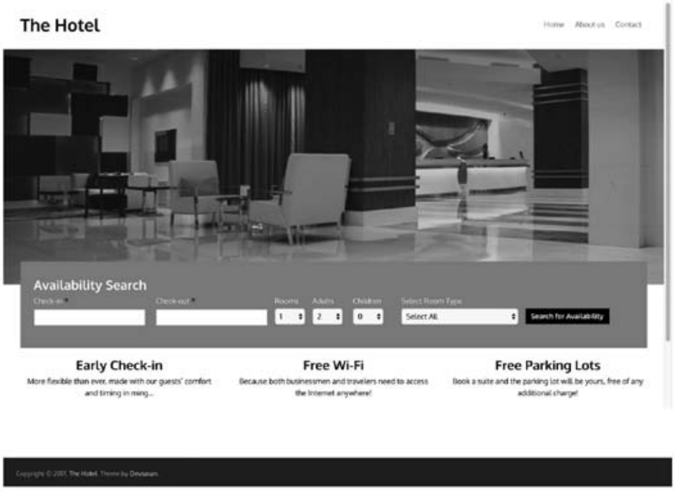
Figure 3 Update messages

3.1.4 Mass customer message updates

Using this functionality, the unit manager can inform a group of customers about a specific situation that affects their bookings. In this way we ensure that all interested customers will get informed even if they do not enter the website for doing that. Similarly, the manager does not have to inform each interested customer one by one.

3.2 Front-End Design of the online application

The basic function of the public section of the application is to demonstrate how the users can, through the search engine, search and book available Hotel rooms. Apart from that, they can also find some information about the services offered by the hotel, the contact form.



Early Check-in
More flexible than ever, made with our guests' comfort and timing in ming...

Free Wi-Fi
Because both businessmen and travelers need to access the Internet anywhere!

Free Parking Lots
Book a suite and the parking lot will be yours, free of any additional charge!

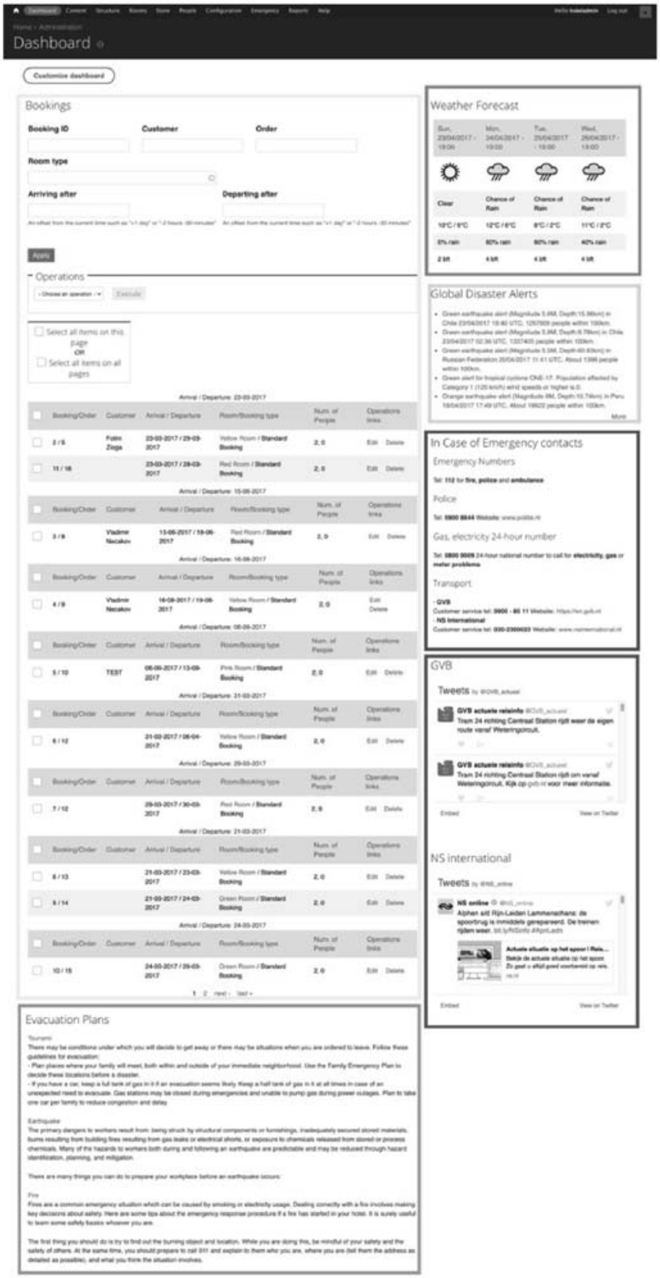
Copyright © 2017, The Hotel. Theme by Devsaran.

Figure 4 Interface for searching and booking accommodation

As we can see in Figure 4, the availability search engine is available only on the homepage of the site. In order to make the reservation, the user is asked to provide his/her personal details and e-mail address (Figure 17). The payment method added to the app is what the hotel accepts as a payment for the duration of the residence. After the user has completed the reservation he/she receives a confirmation message at his/her email address with the details of the reservation.

3.2.1 Bookings Management

At the management of booking part (Figure 5), the manager can inform the customers about any updates regarding their reservation, manage the procedures for Black Swan events, and send mass or individual messages to the customers.



Dashboard

Customize dashboard

Bookings

Booking ID

Customer

Order

Room type

Arriving after

An offset from the current time such as "+1 day" or "-2 hours -30 minutes"

Departing after

An offset from the current time such as "+1 day" or "-2 hours -30 minutes"

Apply

Operations

- Choose an operation - ▾ Execute

□ Select all items on this page
OR
□ Select all items on all pages

Arrival / Departure: 23-03-2017

□	Booking/Order	Customer	Arrival / Departure	Room/Booking type	Num. of People	Operations links
□	2 / 5	Fokin Zioga	23-03-2017 / 29-03-2017	Yellow Room / **Standard Booking**	2, 0	Edit Delete
□	11 / 16		23-03-2017 / 28-03-2017	Red Room / **Standard Booking**	2, 0	Edit Delete

Arrival / Departure: 15-06-2017

□	Booking/Order	Customer	Arrival / Departure	Room/Booking type	Num. of People	Operations links
□	3 / 8	Vladimir Necakov	15-06-2017 / 18-06-2017	Red Room / **Standard Booking**	2, 0	Edit Delete

Arrival / Departure: 16-06-2017

□	Booking/Order	Customer	Arrival / Departure	Room/Booking type	Num. of People	Operations links
□	4 / 9	Vladimir Necakov	16-06-2017 / 19-06-2017	Yellow Room / **Standard Booking**	2, 0	Edit Delete

Arrival / Departure: 06-09-2017

□	Booking/Order	Customer	Arrival / Departure	Room/Booking type	Num. of People	Operations links
□	5 / 10	TEST	06-09-2017 / 13-09-2017	Pink Room / **Standard Booking**	2, 0	Edit Delete

Arrival / Departure: 21-03-2017

□	Booking/Order	Customer	Arrival / Departure	Room/Booking type	Num. of People	Operations links
□	6 / 12		21-03-2017 / 06-04-2017	Yellow Room / **Standard Booking**	2, 0	Edit Delete

Arrival / Departure: 29-03-2017

□	Booking/Order	Customer	Arrival / Departure	Room/Booking type	Num. of People	Operations links
□	7 / 12		29-03-2017 / 30-03-2017	Red Room / **Standard Booking**	2, 0	Edit Delete

Arrival / Departure: 21-03-2017

□	Booking/Order	Customer	Arrival / Departure	Room/Booking type	Num. of People	Operations links
□	8 / 13		21-03-2017 / 23-03-2017	Yellow Room / **Standard Booking**	2, 0	Edit Delete
□	9 / 14		21-03-2017 / 24-03-2017	Green Room / **Standard Booking**	2, 0	Edit Delete

Arrival / Departure: 24-03-2017

□	Booking/Order	Customer	Arrival / Departure	Room/Booking type	Num. of People	Operations links
□	10 / 15		24-03-2017 / 29-03-2017	Green Room / **Standard Booking**	2, 0	Edit Delete

1 2 next › last »

Evacuation Plans

Tsunami

There may be conditions under which you will decide to get away or there may be situations when you are ordered to leave. Follow these guidelines for evacuation:

- Plan places where your family will meet, both within and outside of your immediate neighborhood. Use the Family Emergency Plan to decide these locations before a disaster.

- If you have a car, keep a full tank of gas in it if an evacuation seems likely. Keep a half tank of gas in it at all times in case of an unexpected need to evacuate. Gas stations may be closed during emergencies and unable to pump gas during power outages. Plan to take one car per family to reduce congestion and delay.

Earthquake

The primary dangers to workers result from: being struck by structural components or furnishings, inadequately secured stored materials, burns resulting from building fires resulting from gas leaks or electrical shorts, or exposure to chemicals released from stored or process chemicals. Many of the hazards to workers both during and following an earthquake are predictable and may be reduced through hazard identification, planning, and mitigation.

There are many things you can do to prepare your workplace before an earthquake occurs:

Fire

Fires are a common emergency situation which can be caused by smoking or electricity usage. Dealing correctly with a fire involves making key decisions about safety. Here are some tips about the emergency response procedure if a fire has started in your hotel. It is surely useful to learn some safety basics whoever you are.

The first thing you should do is try to find out the burning object and location. While you are doing this, be mindful of your safety and the safety of others. At the same time, you should prepare to call 911 and explain to them who you are, where you are (tell them the address as detailed as possible), and what you think the situation involves.

Weather Forecast

Sun, 23/04/2017 - 19:00	Mon, 24/04/2017 - 19:00	Tue, 25/04/2017 - 19:00	Wed, 26/04/2017 - 13:00
Clear	Chance of Rain	Chance of Rain	Chance of Rain
10°C / 6°C	12°C / 6°C	8°C / 2°C	11°C / 2°C
0% rain	80% rain	80% rain	40% rain
2 bft	4 bft	4 bft	4 bft

Global Disaster Alerts

- Green earthquake alert (Magnitude 5.4M, Depth:15.96km) in Chile 23/04/2017 18:40 UTC, 1257909 people within 100km.
- Green earthquake alert (Magnitude 5.9M, Depth:8.79km) in Chile 23/04/2017 02:36 UTC, 1337405 people within 100km.
- Green earthquake alert (Magnitude 5.3M, Depth:60.83km) in Russian Federation 20/04/2017 11:41 UTC, About 1396 people within 100km.
- Green alert for tropical cyclone ONE-17. Population affected by Category 1 (120 km/h) wind speeds or higher is 0.
- Orange earthquake alert (Magnitude 6M, Depth:10.74km) in Peru 18/04/2017 17:49 UTC, About 19622 people within 100km.

More

In Case of Emergency contacts

Emergency Numbers

Tel: **112** for **fire, police** and **ambulance**

Police

Tel: **0900 8844** Website: www.politie.nl

Gas, electricity 24-hour number

Tel: **0800 9009** 24-hour national number to call for **electricity, gas** or **meter problems**

Transport

- **GVB**
Customer service tel: **0900 - 80 11** Website: https://en.gvb.nl
- **NS International**
Customer service tel: **030-2300023** Website: www.nsinternational.nl

GVB

Tweets by @GVB_actueel

GVB actuele reisinfo @GVB_actueel
Tram 24 richting Centraal Station rijdt weer de eigen route vanaf Weteringcircuit.

GVB actuele reisinfo @GVB_actueel
Tram 24 richting Centraal Station rijdt om vanaf Weteringcircuit. Kijk op gvb.nl **voor meer informatie.**

Embed View on Twitter

NS international

Tweets by @NS_online

NS online @NS_online
Alphen a/d Rijn-Leiden Lammenschans: de spoorbrug is inmiddels gerepareerd. De treinen rijden weer. bit.ly/NSinfo #AlpnLedn

Actuele situatie op het spoor | Reis...
Bekijk de actuele situatie op het spoor. Zo gaat u altijd goed voorbereid op reis.
ns.nl

Embed View on Twitter

Figure 5 Booking Management Form

4 Conclusion

Black Swan Events or Low Probability High Impact Events (LoPHIEs), like wildfires, earthquakes and volcanic eruptions have significant implications to hospitality and travel industry; a vivid demonstration of the complexity of interconnections between organizational units involved in or affected by LoPHIEs was outlined in this paper through a case study. Appropriate preparations for hospitality and travel SMES to such events have been identified and implemented through a purpose specific information system. To design this platform we followed a requirements analysis methodology. The main elements from the implementation of this platform have also been included whereas elements of an initial usability evaluation have also been presented.

For future development, it is foreseen that novel technological methods and tools that have the potential to handle large volumes of information, fuse heterogeneous data from multiple sensing systems, and take fast decisions at critical turning points must be examined in more detail.

References

1. Hergert, M.: The effect of terrorist attacks on shareholder value: A study of United States international firms. International Journal of Management, Vol. 21 (1), pp. 25—28 (2004).
2. Coombs, W. T., Holladay, J. S.: PR Strategy and Application: Managing Influence, Chichester. Blackwell (2010).
3. Halder, B.: Approaches of Humanitarian Crisis Management - Associated Risks with the ICT-based Crowdsourcing Paradigm (2010). Available at SSRN: http://ssrn.com/abstract=2568233 or http://dx.doi.org/10.2139/ssrn.2568233
4. Coombs, W. T.: Ongoing crisis communication: Planning, managing, and responding, Second Edition, Los Angeles. Sage (2007).
5. Diakou, C.-M., Kokkinaki A. I.: Enabling Sustainable Development through Networks of Collective Intelligence. 4th Conference of the IDRiM Society. Northumbria University, Newcastle. UK, 4-6 September 2013 (2013).
6. Armstrong, J. S.: Findings from evidence-based forecasting: methods for reducing forecast error. International Journal of Forecasting, vol. 22, issue 3, pp. 583-598. (2006).
7. Berg, J. E., Neumann G. R., Rietz, T. A.: Searching for Google's Value: Using Prediction Markets to Forecast Market Capitalization Prior to an Initial Public Offering", Management Science, vol. 55, 3, pp. 348-361 (2009).
8. Goodwin, P., Wright, G.: The limits of forecasting methods in anticipating rare events. Technological Forecasting & Social Change, Vol. 77, pp. 355-368 (2010).
9. Jakoubi, S., Tjoa, S., Quirchmayr, G.: Rope: A Methodology for Enabling the Risk-Aware Modelling and Simulation of Business Processes. ECIS 2007 Proceedings, Paper 47 (2007).
10. Onkal, D., Gonul, M. S.: Judgmental adjustment: a challenge to providers and users of forecasts. Foresight, Vol. 2, pp. 13-17 (2007).
11. Fildes, R., Goodwin, P., Lawrence, M., Nikolopoulos, K.: Effective forecasting and judgmental adjustments: an empirical evaluation and strategies for improvement in supply-chain planning", Int. J. Forecast, Vol. 25, pp. 3-23 (2009).

A Brief Outlining of Entrepreneurship Restructuring as Industrial Waves Occurred

Ioannis Patias

Faculty of Mathematics and Informatics
University of Sofia St.Kliment Ohridski“
5 James Bourchier blvd., 1164, Sofia, Bulgaria
patias@fmi.uni-sofia.bg

Abstract. This paper presents the importance of entrepreneurship restructuring in order to answer the challenges defined by the technological transformations of the industrial revolutions. Technologies are viewed as a phenomenon that can alter both qualities of, but also relationships between, time and space. The purpose of this paper is having on the one side the technologies and their transformations, and on the other the internationalist point of view of entrepreneurship, help in outlining entrepreneurship restructuring as industrial waves occurred. Finally the paper comes to the conclusion that we need to take advantage of our experience gained during the previous industrial revolutions. Meaning that, for sure all industrial revolutions created both wealth, but also entrepreneurship opportunities. But, we also know that they created inequalities. In many cases the technological transformations left people without jobs, and those people often represent those with less abilities to transform. Thus, we need to be proactive and open the discussion on what the entrepreneurship models of the fourth industrial revolution should focus on.

Keywords: globalization, industrialization, entrepreneurship, and industrial revolution

1 Introduction

Many researchers believe that today’s globalized production is not a new phenomenon. Instead they say globalized production is dated back since the early modern days of industrialization. That makes important understanding historically the way that human activity has changed production across time and space. The changes in the production models represent the foundation for the entrepreneurship approach to be used. That makes exploring those models necessary, in order to understand the current state of entrepreneurship restructuring, and later, further the potential restructuring deriving from the application models of the new technologies, like artificial intelligence (AI), three-dimensional (3D) printing, embedded systems, and (Internet of Things) IoT, and how they are affecting, and will affect the applied entrepreneurship models [1].

In this paper, the technological milestones, as they are pre-defined in the entrepreneurship literature, are outlined, described and analyzed according to their effects on existing structures in terms of powers of driving the changes, with focus on the entrepreneurship models.

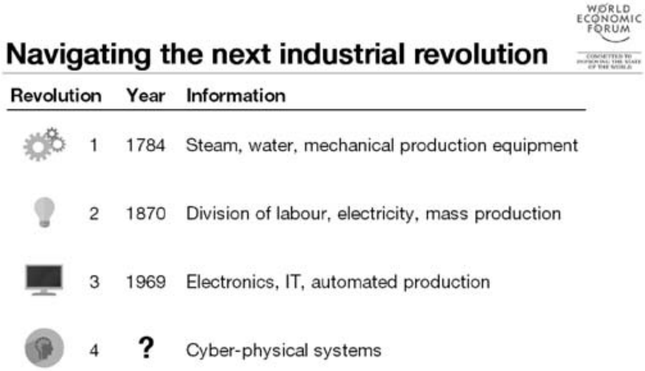
In this review the milestone technologies include all the waves from the steam revolution, electricity, information and communications technology (ICT), containerization and finally the Internet, together with AI, 3D printing, IoT, and embedded systems technologies. For some people Internet can be considered, as a general tool under ICT, but applied here approach of separating ICT from Internet is useful because it makes clear the effect on entrepreneurship restructuring occurring by digitalization and Internet [2].

Central to this paper is thus the notion of technology as a phenomenon that can alter both qualities of, and also relationships between, time and space. So, having on the one side the technologies and their transformations, and on the other the internationalist point of view of entrepreneurship, an outlining of entrepreneurship restructuring as industrial waves occurred is presented.

2. The Revolutions

Technological innovations always accelerated, and so do now, changes in many aspects throughout the global economy. But it is for sure that they always generate both benefits and challenges, and even more this is done in equal measure. In this direction as we are currently witnesses of the fourth industrial revolution, we can see that it is based on three sets of megatrends, which are physical, digital and biological. To get advantage of this this environment, the applied entrepreneurship models should increase, and should be focused on building knowledge and human capital to the benefit of all [3].

But, let’s first have a quick view how we got to the fourth industrial revolutions, by defining the changes occurred during the industrial revolutions. This will help us to see what was the technological transformation, for each wave, and how the entrepreneurship took advantage of this transformation.



Navigating the next industrial revolution

WORLD ECONOMIC FORUM

Revolution	Year	Information
1	1784	Steam, water, mechanical production equipment
2	1870	Division of labour, electricity, mass production
3	1969	Electronics, IT, automated production
4	?	Cyber-physical systems

Figure 1: The industrial revolutions [4]

2.1. The Steam Revolution

The first industrial revolution is known as the steam revolution in the 18th century [5]. The most important change was that by decreasing the transportation costs it was made possible to separate spatially two production pillars, namely processing, and consumption. Out of the model of three production pillars, namely, extraction, processing, consumption, two were separated. The steam opened an entrepreneurship opportunity, which was focused on considering large-scale manufacturing as attractive. This on the other hand further helped and increased this way the international trade. Of course that was made possible, from the institutional point of view, by liberalized regulatory initiatives undertaken internationally.

The separation of the two production pillars was the driver of the changes in the entrepreneurship models. Entrepreneurship needed to resolve two issues, or in other words needed to handle two different sources of driving changes. One is the high cost necessary for coordination and control across space. The second is the localized production, which obviously would be focused on the technologically advanced territories, countries etc.

The changes in the production activities were necessarily related to high levels of localized domestic industrialization. Respectively, at the instructional level, import policies were applied to support building and integration of entire supply chains at home, and keeping this way most of the production activities in-house. That means that the competencies necessary for the production had to be domestic. Or in other words, a competitive country, or nation could be only the one having strong industrial base.

Production globalization was based on the need for both natural resources and new markets to expand. Or in other words the producer had the most important role, and the power, as controller and coordinator of production activities.

2.2. The Electricity Revolution

The second industrial revolution was the one of mass production powered by electricity of the 19-th century. This second wave of technological transformations, apart the production itself was focused also on the progress in communication technology. Communication technology played an important role, by creating networks opening the entrepreneurship opportunity, of making the international production possible on an easier, to coordinate and control across space, manner [6].

But, coordination and control involves managerial skills and technology. Thus, many regions, or countries saw the opportunity, and required the direct transfer of know-how in the form of managerial skills as well as technology.

This was done aiming to protect their local players from the widely opened at international level competition.

2.3. The ICT Revolution

The 20-th century industrial revolution was based on information and communication technologies (ICT). The major change was that ICT further, and this time drastically reduced the control and coordination cost. Although the third industrial revolution is driven by ICT, of course we need also to mention the decreased transports costs occurred by the containerization further helped in this direction. In terms of entrepreneurship opportunity it was clearly defined the outsourcing industry. Outsourcing made manufacturing activities so cheap that production was totally separated and geographically spread. A new wave of trade increase was initiated within supply chains, focused on intermediate products, rather than the typical concerning final products and finished goods [7].

The change of the production activities increased the already existed spatial separation of the production, covering not only those we saw by then, natural resource intensive industries. This trend enabled also the rise of international standards together with the codification of knowledge in order to make the production activities more mobile. And finally by all those trends, this way the entry barriers, typically present to production were decreased.

Now the industry advanced countries focused their efforts in activities like research and development (R&D) and, design, as pre-production activities, and in sales. Meaning R&D as activities related to production, but also sales as the activities following the production, and forming clearly export-oriented industries, in order to take advantage.

The leading role thus was transferred from the producer to the buyer. The entrepreneurship opportunity was on powerful companies, which were established focused on labor-intensive production, and supplying consumer with non-durables. In this direction the leading role went to the hands of the buyer, since he was the one to define the required standards and respectively he was the one to indirectly select the production locations.

2.4. The Internet Revolution

The 21-st century industrial revolution is now focused on Internet production. Starting with the steam revolution where the separation of the production and consumption took place, decreasing the entry barriers to inaccessible until then markets, the ICT revolution further decreased the entry barriers to manufacturing. But, in contrast to the ICT revolution, the Internet revolution made the information distribution both instant and in many cases also free. This helped in making possible for the first time to outsource also intangible activities, which were inappropriate or impossible until then to be outsourced [8].

To have a picture, it is remarkable that by 1993 only 1% of information was shared over the Internet, which reached in 2000, up to 51% and in 2007, got the huge amount of 97% [9, 10]. Comparing with the containerization, which as mentioned supported the increasing trend of goods transportation of the 19th century steam revolution, the Internet revolution is the same for the ongoing increasing trend of intangible activities outsourcing. This of course is also supported by the same requirements for standardization and knowledge codification as was done also in the case of tangible production outsourcing before [11, 12].

The major entrepreneurship opportunity in this period is that we have outsourcing of even more complex activities, like build-on-demand, which derives this standardization, based on technologies like AI, 3D printing, and further enhanced by embedded systems applications, and IoT [13, 14, 15]. Now we have also the entrepreneurship opportunity, having traditional suppliers of goods getting oriented in investing in increasing their brand recognition by increasing customer satisfaction more than ever. But, this task always increases the part of the intangible activities necessary for achieving it.

But, since the intangible activates are also outsourced, that means that we have moved from the concepts of technology transfer, and technology lending to new forms of local technology development. In cheaper markets the approach of reverse engineering created capacities bot for re-design, and development of products in full functionality. Even more, having this technology, the localized improvements can make the final products even more acceptable for the mentioned markets. Adding to this the increased web sales, the initial investor loses any strategic advantage. Of course the buyer gets more power from this transformation, but also there is a huge risk of this the initial investor to loose his interest, and entrepreneurship opportunity in building knowledge and human capital. But, for sure such development will be loose-loose scenario for all. This is why we need to focus on how we can maximize the benefits of science and technology for society [16].

3 Discussion

This paper focuses on bridging two fields, usually discussed separately, those of business and technology. But, it is more than often that business is technology, and in the same time technology is business. Developments in the field of technology are influencing in entrepreneurship opportunities, and visa versa. All these developments, both as technological transformations, and also as entrepreneurship opportunities, occur in time, making changes and transformations an every day standard. Embedded systems technology, AI, IoT, 3D printing, etc. are all well-known technology developments, but as a combination, and the way they are affecting, and will affect entrepreneurship is a field, which worth further research.

History was used as instrument to define the motivation of the changes in entrepreneurship opportunities occurring during the four distinguished technological developments waves, which we also call industrial revolutions. This is because technology developments define the way the production changes. The production changes both during the time, but also in space, as it is transformed from local to global. Beginning with the steam revolution we see how technology influences the cost of transportation of goods, making a globalized approach manageable for the entrepreneurship to deal with. In great similarity we may apply this concept in the 21-st century, fourth industrial revolution of Internet, embedded systems, IoT, AI and 3D printing. We may further try to define how those technology developments can influence, the same way the transportation of intangible goods this time. But, in order to initiate our participation, or to define our entrepreneurship opportunity to a similar with the previous ones, major change, in the business and entrepreneurship we need to use the acquired experience.

All those new activities influence directly the business models applied by the business when developing entrepreneurship initiatives, and opportunities. As the production line concept (extraction, processing, consumption) during the time started from local became global, and we now see it again shifting in new localized manner, the business reacts, identifying entrepreneurship opportunities and develops new models also.

Having this model in mind it is obviously important the use of the technology developments, like the embedded systems technology, IoT, AI, and 3D printing as major factors in the currently applied modern entrepreneurship trends, models and approaches.

5 Conclusions

In brief were presented the four distinguished technological developments waves, called industrial revolutions, together with the changes they brought to the entrepreneurship models applied. We are now in the fourth industrial revolution. Human activities already started changing production across time and space, and having this as a foundation for the entrepreneurship approach to be used, we need to explore the entrepreneurship restructuring. The new technologies driving those changes are almost set. But, the main conclusion we should get is that we need to take advantage of our experience gained during the previous industrial revolutions. Meaning that, for sure all industrial revolutions created both wealth, and also entrepreneurship opportunities. But, we also know that they created inequalities. In many cases the technological transformations left people without jobs, and those people often represented those with less abilities to transform. Thus, we need to be proactive and open the discussion on what the entrepreneurship models of the fourth industrial revolution should focus on.

Acknowledgments. This work is part of the NSF project "GloBIG: A Model of Integration of Cloud Framework for Hybrid Massive Parallelism and its Application for Analysis and Automated Semantic Enhancement of Big Heterogeneous Data Collections", Contract ДН 02/9 of 17.12.2016.

References

1. Encyclopedia Britannica, The Industrial Revolution Economic effects, https://www.britannica.com/topic/history-of-Europe/The-Industrial-Revolution#ref58404, accessed January 2018
2. Special Report, The third industrial revolution, The digitization of manufacturing will transform the way goods are made—and change the politics of jobs too, The Economist, April 22nd 2012
3. Klaus Schwab, The Fourth Industrial Revolution, HBR.ORG, September 2016
4. Nicholas Davis, Member of the Executive Committee, World Economic Forum Geneva, January 2016, https://www.weforum.org/agenda/2016/01/what-is-the-fourth-industrial-revolution/, accessed January 2018
5. Piero Formica, Bologna Shows How a Business Cluster Can Stay Vibrant for Centuries, HBR.ORG, October 2017
6. Marc Levinson, The Industrial Revolution That Never Was, HBR.ORG, July 2014
7. David A. Moss, Confronting the Third Industrial Revolution, HBR.ORG, April 1996
8. Klaus Meyer, and Alexandra Han, Bossard AG: Enabling Industry 4.0 Logistics, Worldwide, HBR.ORG, September 2017
9. Special Report, Ryan Avent, The third great wave, The Economist, October 3rd 2014
10. Hilbert Martin, and Priscila Lopez, The World's Technological Capacity to Store, Communicate, and Compute Information, Science, 2011: 60-65.
11. M. Todorova, D. Orozova, The predicate transformer and its application in introduction to programming courses, Burgas Free University Annual, v. XXII, ISSN: 1311-221-X, 2015, pp. 194-207.
12. M. Todorova, D. Orozova, How to Build up Contemporary Computer Science Specialists – Formal Methods of Verification and Synthsis of Programs in Introduction Courses on Programming, Proceedings of the 9th annual International Conference of Education, Research and Innovation, Seville, 14th, 15th and 16th of November, 2016, Proceedings indexed in WEB of SCIENCE, ISBN: 978-84-617-5895-1, 2016. 4249-4256, doi: 10.21125/iceri.2016.1997, Ref.
13. McKinsey, 3D Printing takes shape, http://www.mckinsey.com/insights/manufacturing/3-d_printing_takes_shape, accessed January 2018
14. Mark Mills, Manufacturing, 3D printing and what china knows about the emerging American century, https://www.forbes.com/sites/markpmills/2011/07/05/manufacturing-3d-printing-and-what-china-knows-about-the-emerging-american-century/#2bb99cdb6f53, accessed January 2018
15. M. Skilton and F. Hovsepian, The 4th Industrial Revolution, https://doi.org/10.1007/978-3-319-62479-2_2, accesses January 2018
16. World Economic Forum, Center for the Fourth Industrial Revolution Agenda, https://www.weforum.org/center-for-the-fourth-industrial-revolution, accessed January 2018

Challenges in Data Anonymity, Protection and Privacy in Medical Research and Services

Anatoliy Dimitrov[1], Dimitar Vassilev[2*]

[1] SAP Labs, Sofia, Bulgaria
[2] Faculty of Mathematics and Informatics, Sofia University "St. Kliment Ohridski",
5 James Bourchier Blvd, Sofia 1164, Bulgaria

* Corresponding author: dimitar.vassilev@fmi.uni-sofia.bg

Abstract. The amount of information gathered and kept world-wide for various purposes grows with alarming rates. The importance of how sensitive and personal information should be handled, has forced legislation changes and new laws to be created for the purpose. Thus, the need for anonymity, data protection and privacy is nowadays not only a moral duty but also a law requirement. This article explains the background and legal framework of data protection. Along with that, programming examples of how such protection can be acquired are given.

Keywords: medical data, anonymity, protection, privacy

1 The importance of anonymity, privacy and the challenges in data protection

Almost every piece of information contains sensitive and private data. Exposing publicly such data could lead to serious financial losses, legal issues and personal inconveniences. That's why data protection and privacy are essential tasks in any data processing project.

When data cannot (public records) or should not (useful information for scientific research) remain private, anonymity or at least pseudonymity must be ensured. Anonymity is derived from the Greek word anonymia meaning "without a name" or "namelessness". The adjective "anonymous" is used to describe situations where the involved person's name is unknown or hidden.

From scientific point of view, the problem with anonymity is that the connection to the original subject may be completely lost. Thus, in some cases important links and dependencies, which could be sometimes very important, will not be revealed. The solution to this challenge is pseudonymity. Pseudonymity is the use of pseudonyms as identifiers [1]. A pseudonym is an identifier of a subject other than one of the subject's real names. On one hand, pseudonymity prevents individuals to be publicly identified and exposed. On the other, it is still

V. Dimitrov, V. Georgiev (Editors): ISGT'2017. ISSN 1314-4855
Proceedings of the 11th International Conference on
Information Systems and Grid Technologies, Sofia, 29.-30. Sep. 2017.

possible, if necessary, to identify them in full compliance with law and privacy requirements.

The choice which data fields are to be protected or anonymised is subjective, but should include all fields that are highly selective, NHS number (in the UK) for example. Less selective fields, such as Birth Date or Postal Code are often also included because they could be cross-matched and lead to a record being identified. Protecting these less identifying fields removes most of their analytic value and should therefore be accompanied by the introduction of new derived and less identifying forms, such as year of birth.

Data fields that are less identifying, such as date of attendance, are usually left untouched. This is mostly because too much statistical utility is lost in doing so. Such an acknowledged and accepted risk is worth in some cases where artificial intelligence will greatly benefit from such genuine information. Unfortunately, such a compromise could lead to the so called "inference attack", e.g. given prior knowledge of a few attendance dates it is possible to identify someone by finding only those people with that pattern of dates. Even more, according to a research "87% (216 million of 248 million) of the population in the United States had reported characteristics that likely made them unique based only on {5-digit ZIP, gender, date of birth}" [2].

Whenever any sensitive data is stored, such as in the case where original data with full personal details must be retained intact, protection against unauthorized access and modification should be implemented. Control access lists should strictly specify personal access rights. Each person with access has to be authenticated in order to verify his identity. This personalized data access control applies not only to the digitally stored data, but also to its physical dimensions.

Last but not least, strong and modern encryption must protect the data so that it is unreadable and unusable even in cases of unauthorized access.

2 Legal regulations and requirements

The importance of privacy protection has been constituted in various legal forms and state regulations around the world. Two of them, in the USA and the EU, have been reviewed below.

In the European Union, the main legal instrument concerning data protection is Directive 95/46/EC1. It defines the protection of individuals with regard to the processing of personal data and on the free movement of such data (Data Protection Directive). Article 8 of this directive specifies special categories of data that shouldn't be processed. One of these categories is data concerning health.

In the US, there is a large number of data privacy regulations on state level but on federal level the most important ones are The Federal Trade Commission Act (15 U.S.C. §§41-58) (FTC Act) and the Health Insurance Portability and

Accountability Act (HIPAA) (42 U.S.C. §1301 et seq.). The latter is especially important because it regulates the use of medical information as in cases of scientific researches.

In essence, both in the European Union and in the United States, consent is required for the use of personal information. However, this does not apply if the data has been anonymised and individuals cannot be identified through linking the information to other publicly available data. [3,4].

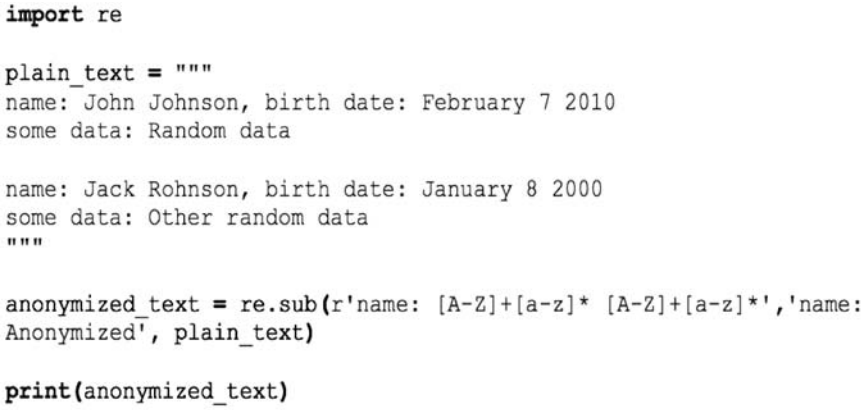
3 Methods for ensuring anonymity, data protection and privacy

There are complete solutions, both commercial and open source, for the protection of data and privacy. Popular data anonymization tools such as ARX (http://arx.deidentifier.org/) have advanced features to anonymize any sensitive personal data. Furthermore, strong encryption algorithms are supported out of the box by every modern computer operating system including Windows, Mac and Linux based such. There are also third party encryption solutions such as the Encryption Wizard developed and used by the American Army and Air Force (https://www.spi.dod.mil/ewizard.htm).

If a custom solution is needed for data protection, there are plenty of readily available libraries and modules in every popular programming language. For example, Python supports fast and powerful implementation both of anonymization and encryption as the next examples show.

3.1. Anonymization and pseudonymization

For a simple anonymization, regular expressions with search and replace functions can be used. Here is an example:



```
import re

plain_text = """
name: John Johnson, birth date: February 7 2010
some data: Random data

name: Jack Rohnson, birth date: January 8 2000
some data: Other random data
"""

anonymized_text = re.sub(r'name: [A-Z]+[a-z]* [A-Z]+[a-z]*','name:
Anonymized', plain_text)

print(anonymized_text)
```

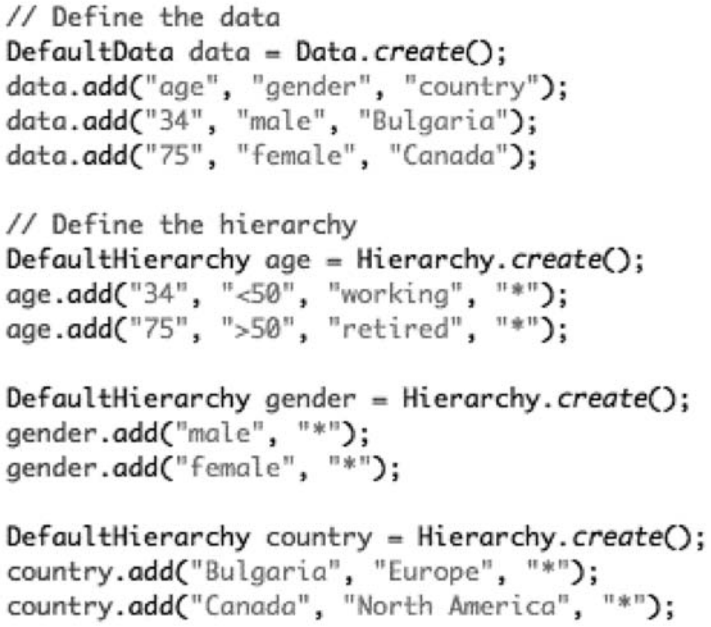
The above Python code will accomplish anonymization by replacing every occurrence of a name in the form of two alphabetical words with initial capital letter following the string “name:”.

The above code is written specifically for the text example above and it can be further enhanced and customized to specific needs with different order of names and personal details. That’s thanks to the powerful and flexible regular expressions, supported in Python and many other programming languages.

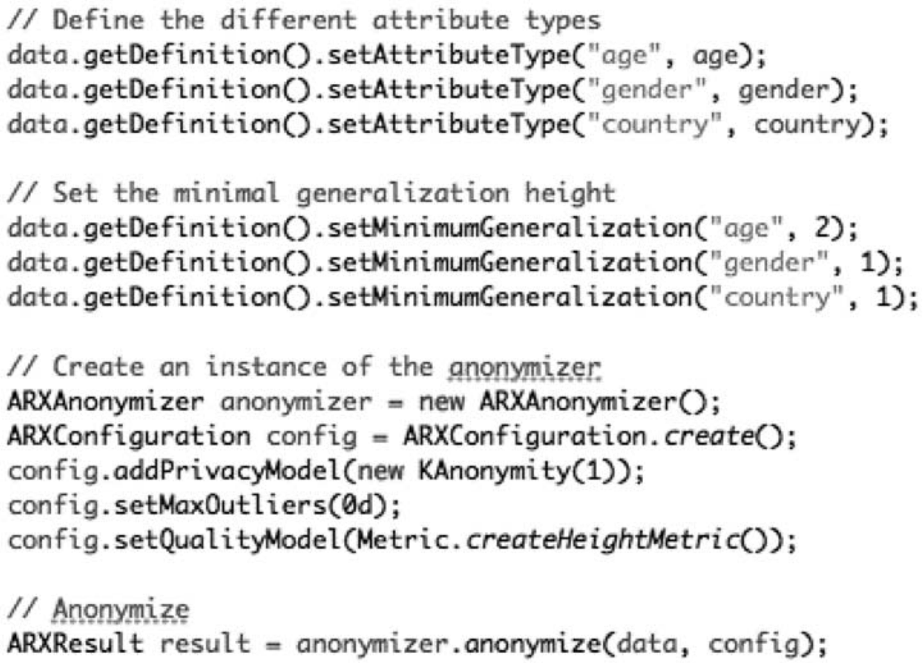
Names, and other personal details, can be also replaced with unique pseudonyms, for achieving pseudonymity. In programming, e.g. Python or Java, this can be done using a dictionary and inserting a unique key – value pair for every name replaced by an automatically generated pseudonym. In this approach it will be essential from security point of view that this newly created dictionary is stored separately from the main file in a secure manner

3.2. Generalization

In some cases, generalization is preferred than pure anonymization because certain statistical relations are important and should not be lost. For example, you can generalize the location of a subject from city to state. Thus, with sacrificing some of the precision you are still able to reveal a geographical distribution while not compromising the privacy of the individual. Here is an example for generalization with Java and the Arx library aforementioned:



```
// Define the data
DefaultData data = Data.create();
data.add("age", "gender", "country");
data.add("34", "male", "Bulgaria");
data.add("75", "female", "Canada");

// Define the hierarchy
DefaultHierarchy age = Hierarchy.create();
age.add("34", "<50", "working", "*");
age.add("75", ">50", "retired", "*");

DefaultHierarchy gender = Hierarchy.create();
gender.add("male", "*");
gender.add("female", "*");

DefaultHierarchy country = Hierarchy.create();
country.add("Bulgaria", "Europe", "*");
country.add("Canada", "North America", "*");
```



```
// Define the different attribute types
data.getDefinition().setAttributeType("age", age);
data.getDefinition().setAttributeType("gender", gender);
data.getDefinition().setAttributeType("country", country);

// Set the minimal generalization height
data.getDefinition().setMinimumGeneralization("age", 2);
data.getDefinition().setMinimumGeneralization("gender", 1);
data.getDefinition().setMinimumGeneralization("country", 1);

// Create an instance of the anonymizer
ARXAnonymizer anonymizer = new ARXAnonymizer();
ARXConfiguration config = ARXConfiguration.create();
config.addPrivacyModel(new KAnonymity(1));
config.setMaxOutliers(0d);
config.setQualityModel(Metric.createHeightMetric());

// Anonymize
ARXResult result = anonymizer.anonymize(data, config);
```

3.3. Authentication and authorization

Authentication represents the process by which one subject verifies the identity of another, and must be performed in a secure fashion; otherwise a perpetrator may impersonate others to gain access to a system. Authentication typically involves the subject demonstrating some form of evidence to prove its identity [5].

Once authentication has successfully completed, access controls should be enforced upon the principals associated with the authenticated subject. The more detailed the access controls are, the better the data protection will be. As a rule of thumb, as little permissions should be granted by default.

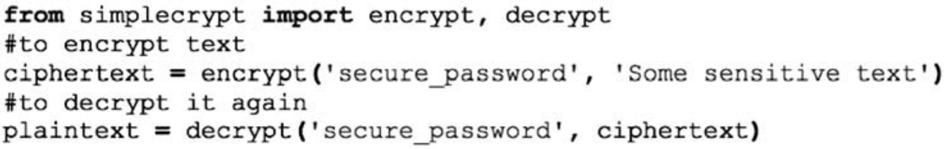
3.4. Encryption

No sensitive information should not be kept in clear, readable text format at any time. Instead, it must be encrypted so that it cannot be understood, nor exploited in case of an unauthorized access.

The process of transforming plaintext into ciphertext is called encipherment or **encryption**. A cipher is a secret method of writing, whereby plaintext (or cleartext) is transformed into ciphertext. [6]

The reliability of an encryption method is determined by the strength of its algorithm and the length of its key. At the writing of this document, the algorithm called Advanced Encryption Standard (AES) with 256 bit key length is the most widely deployed and accepted method for encryption. It renders 2^{256} combinations that have to be broken in order the encrypted document to be compromised. With the current computer power it is unfeasible to pursuit an approach of brute-force finding the correct combination. However, it should be noted that in the past, weaknesses of previously popular encryption algorithms have allowed an attacker to decrypt the information very fast without having to go through the combinations.

Thanks to its popularity, AES is widely supported in programming and every modern language has implementation for it. Python has a library called PyCrypto (https://www.dlitz.net/software/pycrypto/) which supports AES with 256 bit key length. The use of this library has been further facilitated by additional modules such as Simple-crypt (https://pypi.python.org/pypi/simple-crypt) which make it very easy and simple to use encryption. Here is an example:



```
from simplecrypt import encrypt, decrypt
#to encrypt text
ciphertext = encrypt('secure_password', 'Some sensitive text')
#to decrypt it again
plaintext = decrypt('secure_password', ciphertext)
```

This ease of use of encryption further encourages its use and makes it possible to be applied in wider areas.

4 Conclusion

Anonymity, data protection and privacy are more than just an essential in today’s informational age. Whether for business or for scientific needs, the collected and stored data must comply with the corresponding data protection standards and official regulations. This has led to the development of many third-party tools and programming languages libraries which make the implementation of data protection mechanisms simple and easy.

5 Acknowledgements:

The presented work has been funded by the Bulgarian NSF within the “Methods for Data Analysis and Knowledge Discovery in Big Sequencing Datasets” project, Contract DFNI-I02/7 of 12 December 2014.

6 References

[1] Andreas Pfitzmann, M. H. (2008). Anonymity, Unlinkability, Undetectability, Unobservability, Pseudonymity, and Identity Management – A Consolidated Proposal for Terminology.

[2] Sweeney, L. (2000). Simple Demographics Often Identify People Uniquely. Pittsburgh: Carnegie Mellon University.

[3] The European Parliament and of the Council of Europe. (1995). Directive 95/46/EC. Official Journal of the European Communities 1995.

[4] US Government Printing Office. (1996). Health insurance portability and accountability act. US Government Printing Office

[5] Charlie Lai Li Gong, L. K. (n.d.). User Authentication and Authorization in the JavaTM Platform.

[6] Dorothy Elizabeth Rob, l. D. (1982). Cryptography and Data Security

Integrated Semantic Data Model for Functional Annotation of Protein Sequences

Deyan Peychev[1*], Dimitar Vassilev[2]

[1] AgroBioInstitute, 8 Dragan Tsankov Str., Sofia 1164, Bulgaria
[2] Faculty of Mathematics and Informatics, Sofia University "St. Kliment Ohridski",
5 James Bourchier Blvd, Sofia 1164, Bulgaria
* Corresponding author: deyan.pey@gmail.com

Abstract. Functional annotation of proteins is a key phase in the large process of analyzing de-novo sequenced genomes. Often software annotation tools are insensitive to eliminate error annotations that are associated with controversy and inconsistency in biological aspect. In this study we introduce semantic web model to represent functional annotations based on Resource Description Framework (RDF) standard. We integrate several databases with protein sequence information and ontologies describing functional relationships of protein molecules. Using inference based on Web Ontology Language (OWL), RDF database engines are able to take decisions which co-occurred candidate annotations should be marked as inconsistent if they don't pass the reality checks related to function's co-existence, subcellular location and species affiliation [1]. This approach reduces the number of false positives and time spend for quality checks of machine annotations. Introduced data model is designed to combine semantic representation of annotations with data samples intended for machine learning, which is the common way to generate machine annotations. Current work is part of large scale project of functional annotation of plant genomes.

Keywords: functional annotation, semantic web, data model, machine learning.

1 Introduction

A number of software tools are available to deal with challenges of NGS data output. They are intended for automation of most of the steps involved – such as error detection, assembly process, gene prediction and annotation of resulting sequences. Genome annotation process is limitation step of entire sequencing workflow due to difficulties in searching and interpretation of different candidate reference gene records available in biological databases. Due to the lack of standards for knowledge representation in the field of sequencing analysis, the major problems are related mostly to semantics of annotation description and quality checks rather than finding possible matches between sequenced fragments and gene records in databases. Furthermore, expert evaluation of automated annotation shows that most of the problems of annotation performance come from inconsistencies such as over annotation or fail of basic reality checks, which

V. Dimitrov, V. Georgiev (Editors): ISGT'2017. ISSN 1314-4855
Proceedings of the 11th International Conference on
INFORMATION SYSTEMS AND GRID TECHNOLOGIES, Sofia, 29.-30. Sep. 2017.

increases false positives in assigning molecular functions to sequenced regions. Thus assigning of gene functions appears to be the bottleneck of NGS data analysis and its performance highly depends of relatively slow process of manual evaluation and curating to achieve correct knowledge about molecular functions of de-novo sequenced regions [2].

The goal of current work is design and implementation of semantic model for formal description of some protein sequence databases, to provide integrated repository of input samples for machine learning process. Also we experiment with possibilities to track inconsistencies between machine annotations just using capabilities of OWL Lite reasoning and class restrictions in Gene Ontology [3]. Reducing of inconsistencies will also reduce the effort and time spent to evaluate annotation quality.

We use RDF standard to represent referenced database records and their relationships to controlled vocabularies, families, patterns, motifs and more entities used in the process of annotation. Formal description of such entities and relationship provides huge possibilities for knowledge extraction and representation of implicit machine conclusions, which further can be used to automate large number of checks performed by curators [4]. Also we use a service API for easy navigation and querying of resulting network based on SPARQL protocol. With SPARQL endpoints researchers are capable to link result records from different scientific groups together in attempt to improve performance and quality of the process of de-novo sequencing of multiple genomes.

2 Materials and methods

Several public datasets with protein information were integrated into the semantic model using common identifiers as references. These datasets contain information needed to achieve both - data used as patterns to build machine learning models and data used as labels for prediction of protein function. Also the semantic model allows machine annotations to be stored and automatically marked if they are inconsistent with the restrictions of the ontology which defines class definitions and semantics of the relationships of predicted labels.

UniPriot - major protein resource of protein sequence information that contains protein sequences, features and lot of human annotations about protein functions. Actual part of the information from UniProt is:

1. Protein sequences from multiple species which are annotated by humans only but not by machines and has references to PROSITE database.
1. Manual annotated records as direct relations to Gene Ontology terms [5].

PROSITE - database of protein patterns of domains, families and functional motifs containing detailed information about literature citations where the information about pattern composition is obtained from. PROSITE motif descriptions and literature citations are also included in machine predicted annotations. Extracted information contains:

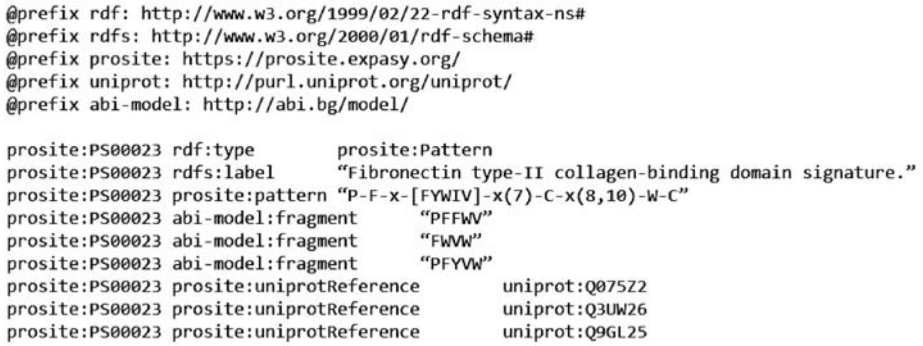
1. Patterns - regex-like representation of the rules which are applied to compose pattern that matches a set of proteins.
1. Cross-references to UniProt entries. Every pattern is linked to set of UniProt entries.
2. Literature citations linked to every pattern [6].

Gene Ontology - major resource for protein annotation which describes proteins in the terms of function, subcellular location and the biological process in which the protein is involved. In current study the disjoint restrictions of Gene Ontology are used to define semantics of reality checks applied in entire semantic model. The usage of "disjointWith" OWL property from Gene Ontology is main property applied to discard false positive annotations generated by machine learning prediction process. In current study only molecular function branch of the terms is included with all associated information from the ontology.

Interpro - protein families and subfamilies are linked to UniProt protein resources. Interpro is used to achieve additional info about protein domains and motifs. It also contains machine annotations to Gene Ontology terms that can be used as benchmark to compare different machine learning annotation methods [7].

2.1. Data preparation

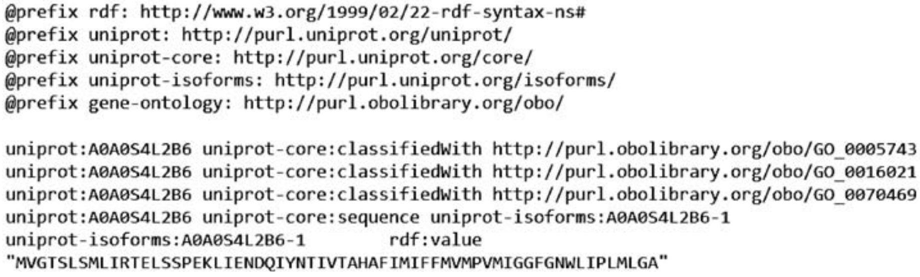
PROSITE patterns are regex-like and they are linked to UniProt sequences that contains matches for them. The first transformation task is to convert the patterns to real regex patterns. Thus they are capable to be executed against linked corresponding sequences and to retrieve actual matching amino acid fragments that characterize the link between the pattern and protein sequence. These amino acid fragments later can be used to train machine learning algorithms such as Support Vector Machine or J48 in order to predict which Gene Ontology term describe possible molecular function of de-novo sequenced proteins. PROSITE general entry information, generated amino acid fragments and cross-referenced UniProt identifiers are converted to RDF format keeping all persistent identifiers of the original data sources intact. The RDF schema is designed to keep all resource URIs resolvable where it is possible for quick navigation to original sources. Following basic schemata of OWL classes and properties, this transformation can be described in N-Triples-like syntax as follows [fig. 1]:



```
@prefix rdf: http://www.w3.org/1999/02/22-rdf-syntax-ns#
@prefix rdfs: http://www.w3.org/2000/01/rdf-schema#
@prefix prosite: https://prosite.expasy.org/
@prefix uniprot: http://purl.uniprot.org/uniprot/
@prefix abi-model: http://abi.bg/model/

prosite:PS00023 rdf:type        prosite:Pattern
prosite:PS00023 rdfs:label      “Fibronectin type-II collagen-binding domain signature.”
prosite:PS00023 prosite:pattern “P-F-x-[FYWIV]-x(7)-C-x(8,10)-W-C”
prosite:PS00023 abi-model:fragment      “PFFWV”
prosite:PS00023 abi-model:fragment      “FWVW”
prosite:PS00023 abi-model:fragment      “PFYVW”
prosite:PS00023 prosite:uniprotReference        uniprot:Q075Z2
prosite:PS00023 prosite:uniprotReference        uniprot:Q3UW26
prosite:PS00023 prosite:uniprotReference        uniprot:Q9GL25
```

Figure 1 Transformation syntax

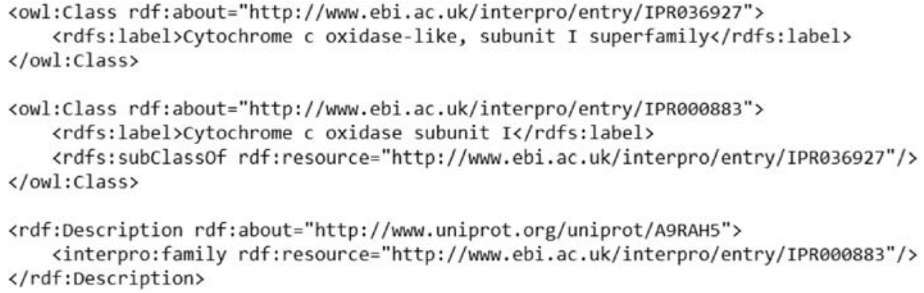
UniProt and Gene Ontology already have their own RDF distributions and can be directly referenced with persistent resolvable URIs. UniProt is huge information resource and that is the reason to use only part of the database related to protein sequences and Gene Ontology annotations. Because of persistent URIs used to describe the resources, these data sets can be directly mapped to PROSITE RDF serialization without any modifications. N-Triples-like syntax as follows:



```
@prefix rdf: http://www.w3.org/1999/02/22-rdf-syntax-ns#
@prefix uniprot: http://purl.uniprot.org/uniprot/
@prefix uniprot-core: http://purl.uniprot.org/core/
@prefix uniprot-isoforms: http://purl.uniprot.org/isoforms/
@prefix gene-ontology: http://purl.obolibrary.org/obo/

uniprot:A0A0S4L2B6 uniprot-core:classifiedWith http://purl.obolibrary.org/obo/GO_0005743
uniprot:A0A0S4L2B6 uniprot-core:classifiedWith http://purl.obolibrary.org/obo/GO_0016021
uniprot:A0A0S4L2B6 uniprot-core:classifiedWith http://purl.obolibrary.org/obo/GO_0070469
uniprot:A0A0S4L2B6 uniprot-core:sequence uniprot-isoforms:A0A0S4L2B6-1
uniprot-isoforms:A0A0S4L2B6-1        rdf:value
"MVGTSLSMLIRTELSSPEKLIENDQIYNTIVTAHAFIMIFFMVMPVMIGGFGNWLIPLMLGA"
```

Figure 2 Serialization syntax

Interpro hierarchy of protein families is converted to RDF format from its’ XML distribution using rdfs:subClassOf property. UniProt identifiers that refer to protein families are converted to be URIs with “http://purl.uniprot.org/uniprot/” prefix to fit directly to the rest of the schema. With RDF-XML syntax this can be described as follows:

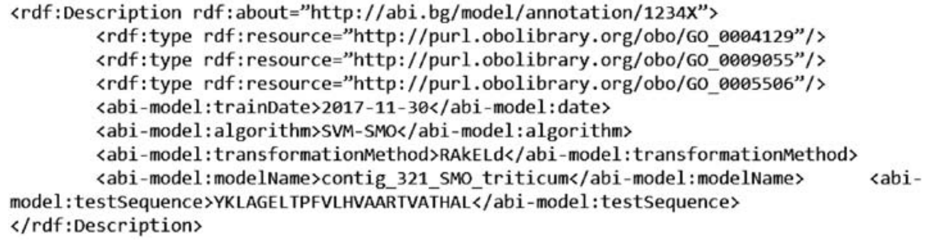


```
<owl:Class rdf:about="http://www.ebi.ac.uk/interpro/entry/IPR036927">
    <rdfs:label>Cytochrome c oxidase-like, subunit I superfamily</rdfs:label>
</owl:Class>

<owl:Class rdf:about="http://www.ebi.ac.uk/interpro/entry/IPR000883">
    <rdfs:label>Cytochrome c oxidase subunit I</rdfs:label>
    <rdfs:subClassOf rdf:resource="http://www.ebi.ac.uk/interpro/entry/IPR036927"/>
</owl:Class>

<rdf:Description rdf:about="http://www.uniprot.org/uniprot/A9RAH5">
    <interpro:family rdf:resource="http://www.ebi.ac.uk/interpro/entry/IPR000883"/>
</rdf:Description>
```

Figure 3 RDF-XML syntax

2.2. Annotation schema

Data sources mentioned in previous section are transformed to RDF format and linked together with common persistent identifiers in the form of URIs. This part of the semantic model is suitable to feed instances required to train machine learning model. One more schemata is designed to describe and store predictions from machine learning model in order to provide semantic functional annotations about new proteins which are unseen by machine learning model. This schema contains amino acid sequence which is analyzed, predicted Gene Ontology terms linked as labels and some metadata for details of the analysis - such as the algorithm used for classification, the name of the problem transformation method used if the classification task is considered to be multi-label task, date of the analysis and the name of the training dataset. To make the link between annotation object and particular Gene Ontology class, the build-in RDF property “rdf:type” is used. Example of predicted annotation resource can be described as follows:



```
<rdf:Description rdf:about=”http://abi.bg/model/annotation/1234X”>
        <rdf:type rdf:resource=”http://purl.obolibrary.org/obo/GO_0004129”/>
        <rdf:type rdf:resource=”http://purl.obolibrary.org/obo/GO_0009055”/>
        <rdf:type rdf:resource=”http://purl.obolibrary.org/obo/GO_0005506”/>
        <abi-model:trainDate>2017-11-30</abi-model:date>
        <abi-model:algorithm>SVM-SMO</abi-model:algorithm>
        <abi-model:transformationMethod>RAkELd</abi-model:transformationMethod>
        <abi-model:modelName>contig_321_SMO_triticum</abi-model:modelName>      <abi-
model:testSequence>YKLAGELTPFVLHVAARTVATHAL</abi-model:testSequence>
</rdf:Description>
```

Figure 4 Predicted annotation resource description

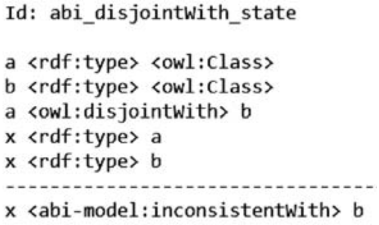
2.2. Storage engine and inference rules

After RDF transformations, data sets are loaded to RDF-triple storage engine “GraphDB” provided by Ontotext AD (http://graphdb.ontotext.com). Inference applied is based on one of the predefined GraphDB reasoners “OWL-Horst” and axioms defined in Gene Ontology and Interpro class hierarchy. Gr aphDB also supports custom inference rules which generates custom extensions of the default reasoner. The syntax of the rules is:

<premis> [Optional constraint]

<consequence> [Optional constraint]

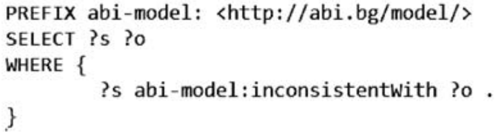
One custom rule is defined to follow disjointness between classes from Gene Ontology in order to materialize inconsistencies derived from machine annotations. Thus semantic model provide possibilities to improve the machine learning model prior every train iteration:



```
Id: abi_disjointWith_state

a <rdf:type> <owl:Class>
b <rdf:type> <owl:Class>
a <owl:disjointWith> b
x <rdf:type> a
x <rdf:type> b
--------------------------------
x <abi-model:inconsistentWith> b
```

Figure 5 Improving machine learning model

This rule at inference time states the following logic - if there are two classes “a” and “b” and they are in disjoint, then for every annotation instance which is of type of both classes additional implicit statement will be generated indicating inconsistent link between the annotation object and one of both classes. This is convenient way to analyse the number and the nature of inconsistencies just using SPARQL queries like:



```
PREFIX abi-model: <http://abi.bg/model/>
SELECT ?s ?o
WHERE {
        ?s abi-model:inconsistentWith ?o .
}
```

Figure 6 SPARQL query example

For annotation example mentioned above in the article, the result of this query should be the following:

http://abi.bg/model/annotation/1234X | http://purl.obolibrary.org/obo/GO_0005506

Unlike OWL DL, OWL Lite does not allow the use of owl:disjointWith [8].

Using OWL Lite reasoning the usage of owl:disjointWith can be interpreted as not built-in property and consistency checks will never fail, but at the same time we can easily trigger disjointness applying custom inference rule to generate implicit statements to materialize inconsistencies as regular RDF statements.

Detailed documentation about GraphDB inference is available at:

http://graphdb.ontotext.com/documentation/standard/reasoning.html

Custom Java classes are developed and used to preprocess the data from different sources. They include parsers for the distribution formats, pattern regex converter, regex matcher and RDF syntax generator.

Protein sequences from UniProt are converted to plain capitalized strings for easier processing and UniProt identifiers are matched between both PROSITE and UniProt databases using cross-reference links.

MySQL database is used as intermediate store to organize the data prior RDF statements generation. Two tables are defined to store relationships between extracted patterns and the sequences in which they occurred.

InterPro hierarchies are transformed to RDF from its' XML distribution format using XSLT transformation. All generated RDF data is saved in RDF/XML or Ntripple files.

After all transformations RDF data is loaded in semantic repository GraphDB v8.3 using "Horst" flavor of OWL Lite language.

SPARQL queries are executed inside GraphDB workbench or directly via SPARQL endpoint.

3 Results and discussion

PROSITE database is represented with 2511 unique patterns in release 2017_11 and all of them are processed as amino acid fragments linked to corresponding sequences from UniProt. Total number of 268.692 unique relationships between all data sources are loaded in semantic repository for two hours without inference. Later the inference over loaded triples took 16 hours to infer OWL Lite axioms over Gene Ontology classes on 8 CPU core machine with 32G of RAM. Overall number of statements - both explicit and implicit after the inference step is scaled to 2.5 million statements.

To test the inference of the semantic model, one batch of 300 simulated instances of machine annotations are loaded with total number of 32 inconsistencies found and materialized as implicit RDF statements. Simulated annotations are generated with the principle to add some Gene Ontology terms which are in disjoint as classes for one annotation. For example the terms GO:0003690 and GO:0003697 which are actually "double-stranded DNA binding" and "single-stranded DNA binding" terms. They are in disjoint as Gene Ontology classes and didn't pass the reality check in biological aspect because two terms are opposite

[1]. The special predicate “abi-model:inconsistentWith” is implicitly generated between annotation node and the second term linked to that node (in this case GO:0003697).

The opened question is which Gene Ontology term do not pass the reality check. Actually both classes has explicit disjointWith statements for each other and just the order of the appearance is necessary but not sufficient criteria to choose one of them. One probable solution is to keep the term with higher probability score generated by machine learning algorithm at prediction time. This is possible if the classification task is defined as multi-label task and probabilities distribution of labels is returned for every prediction. Descending order of Gene Ontology terms thus will provide possibility first to add the term with highest score into semantic repository, next is the second one etc. If some of the terms are in disjoint with other terms from label collection which are already added to the repository, then this term will be marked as inconsistent.

4 Conclusion

Semantic technologies can have a major impact on the quality of machine learning strategies, which takes an increasing role in the process of annotating newly-sequenced genomes. Despite of the quality of the predictions made by machines, quality check by curators is mandatory task in order to track and discard logical inconsistencies of prediction in biological aspect. Semantic networks are capable to deal with human knowledge in machine understandable way providing great possibilities for researchers to encode domain specific knowledge about properties of objects of interest like genes and proteins.

Often curation process is time consuming effort to validate machine learning models and quality of annotations at all. Using knowledge stored in ontologies, semantic networks can reduce probabilities of false positive annotations, which are result of inconsistencies between proposed concepts and conclusions of learned models. Future work in the field of current study is to apply machine learning annotations from real use case scenario and evaluate them through manually annotated dataset.

5 Acknowledgements:

The presented work has been funded by the Bulgarian NSF within the “GloBIG: A Model of Integration of Cloud Framework for Hybrid Massive Parallelism and its Application for Analysis and Automated Semantic Enhancement of Big Heterogeneous Data Collections” project, Contract ДН 02/9 of 17.12.2016.

6 References

1. Koonin EV, Galperin MY (2003) Sequence - Evolution - Function: Computational Approaches in Comparative Genomics. Boston: Kluwer Academic
2. Poux S, Arighi, CN, Magrane M, Bateman A, Wei C-H, Lu Z, Boutet E, Bye-A-Jee H (2017) On expert curation and sustainability: UniProtKB/Swiss-Prot as a case study bioRxiv
3. The Gene Ontology Consortium (2013) Gene Ontology Annotations and Resources. Nucleic Acids Research, Volume 41, Issue D1, 1, Pages D530–D535
4. Claude Pasquier, Biological data integration using Semantic Web technologies. Institute of Signaling, Developmental Biology & Cancer, CNRS - UMR 6543, University of Nice Sophia-Antipolis
5. Pundir S, Martin MJ, O’Donovan C (2017) UniProt Protein Knowledgebase. Methods Mol. Biol. 1558:41-55
6. Christian J. A. Sigrist, Lorenzo Cerutti, Edouard de Castro, Petra S (2010) PROSITE, a protein domain database for functional characterization and annotation. Nucleic Acids Res. 2010 Jan
7. R. Apweiler T. K. Attwood A. Bairoch A. Bateman E. Birney M. Biswas P. Bucher L. (2001) The InterPro database, an integrated documentation resource for protein families, domains and functional sites. Nucleic Acids Research, Volume 29, Issue 1, Pages 37–40
8. W3C Recommendation 10 February 2004; OWL Web Ontology Language Reference

Discovery of Pattern Deviations Caused by Gene Duplication in Plant Evolutionary Trees

Irena Avdjieva*, Milko Krachunov, Ognyan Kulev, Dimitar Vassilev

Faculty of Mathematics and Informatics, Sofia University "St. Kliment Ohridski",
5 James Bourchier blvd., Sofia 1164, Bulgaria
* Corresponding author: i.y.avdjieva@fmi.uni-sofia.bg

Abstract. The purpose of the current work is to develop a pipeline for recognizing patterns in evolutionary tree nodes caused by gene duplication. Such duplications result in deviation from the typical ratio between nodes representing genes of interest. These genes are supposed to be involved in the evolution of the economically significant drought-resistant C4 photosynthesis modification in plants. To predict those genes, the study uses a large dataset of plant phylogenetic trees and is focused on four cereals – two C3 (rice and brachypodium), and two C4 plants (sorghum and maize). The pipeline utilizes the tree-reading abilities of the ETE toolkit framework and is written in Python 2.7. It involves several stages of processing the dataset: finding objects of interest, defining the expected ratio between nodes, detecting deviations and as a result recognizes and isolates the nodes causing them for additional analyses and verification.

Keywords: evolution, pattern recognition, homology, Python programming

1 Introduction

In biology, the evolutionary relationships between entitites (genes, species, populations or other operational taxonomic units - OTUs) are graphically represented as a phylogenetic tree. A phylogenetic tree is an onedirectional bifurcating graph where end nodes represent the OTUs, internal nodes are common ancestors and branching indicates events of duplication or speciation. Thus, all OTUs in one tree are called a family because they have a common origin. Comparative phylogenetic allows the gain, loss and evolution of genes associated with important traits to be traced.

In this study, we use computational methods to analyze plant phylogenetic trees in order to predict genes involved in the evolution of C4 photosynthesis – a modification that allows plants to minimize water loss and utilize atmospheric CO_2 more efficiently in warm and dry conditions.

1. 1. Popular hypotheses for the evolution of C4 genes

Research shows that C4 photosynthesis has evolved independently more than 60 times during the evolution of plants and is currently observed in about 3%

V. Dimitrov, V. Georgiev (Editors): ISGT'2017. ISSN 1314-4855
Proceedings of the 11th International Conference on
Information Systems and Grid Technologies, Sofia, 29.-30. Sep. 2017.

of known plant species [1]. There are several hypotheses about the development of this modification:1) genes that are present in C4 plants but not in C3 plants (or *vice versa*); 2) genes that are duplicated in C4 plants but are present as single copies in C3 plants (or *vice versa*), and 3) copy number variations between homologous genes in one of the two groups.

When taking into account recent research on C4 plants, the "duplicated vs. single copies" hypothesis is deemed most accurate. To predict genes that may be involved in C4 photosynthesis this study proposes a comparative phylogenetic approach that involves finding a certain pattern in the topology of the trees which contain genes form both C3 and C4 species. Using phylogenetic trees as source data saves the need to do sequence analyses and covers a large part of the genomes of the species involved in this study.

2 Materials and methods

In order to find suitable candidate-genes for our study, we started with a large dataset of plant gene trees from different species. We chose to work with publicly available phylogenetic data from the database Ensembl Plants [2]. It contains more than 40 species – mostly model and/or economically important plants. The dataset consists more than 100 000 gene tress and had to undergo several stages of filtering by varions criteria so that less than 100 genes would be proposed as candidates for involvement in C4 photosynthesis.

The input data is an emf flatfile dump containing phylogenetic trees in Newick format and additional data for each gene in the corresponding tree. The individual entries were separated by two vertical slashes (//) and the tree format was separated from the additional data with a single line (DATA).

2.1. Establishing a pattern

Before proceeding to search for genes potentially involved in C4 photosynthesis, it was necessary to select appropriate objects. Since most C4 plants are cereals, he study was focused on this group and two representatives from the two photosynthetic groups were chosen: C3 plants rice (Oryza sativa) and Brachypodium distachyon, and C4 plants maize (Zea mays) and sorghum (Sorghum bicolor). The reason for choosing exactly these four species to study the evolution of C4 photosynthesis is justified not only by the fact that they belong to the same family, but mainly because they are subject to intense study because of their extremely important economic importance. According to FAOSTAT [3], rice, maize and sorghum are ranked respectively in first, third and fifth places of world-wide cereals, and the brachypodium is a model plant for all wheat [4]. Therefore, their genomes are better annotated than those of a number of other plants.

The typical ratio of the genes of these four species in the trees is 1: 1: 1: 2 [5], as maize undergoes a total genomic duplication dating back to 5-12 million years after the maize and sorghum have been separated as individual species.

2.2. Dataset preparation and pattern-finding script

The first stage of this process begins by filtering the tree array and work continues only with those in which the species under consideration are represented by at least one gene. The rest of the task was broken down into several stages:

Stage 1: Isolation of duplicated genes from the least possible subsurface where a deviation from the standard 1: 1: 1: 2 ratio is observed;

Stage 2: Tracking and Removing Repetitive Genes;

Stage 3: Statistical analysis of the distribution of genes and species in isolated trees;

Stage 4: Calculating the distance between duplicated genes;

Stage 5: Determining the threshold value of the distance;

Stage 6: Error correction and classification overrides in some groups of genes.

The next reduction step is to search trees for duplicated genes. This means that the ratio of 1: 1: 1: 2 for the four species in such trees will be violated. This can be ascertained by simply counting the genes, but it would only produce a true result if all four species are located in just one well. If this condition is not fulfilled, there is a risk of misinterpretation: the ratio in individual wells is not the same as standard, but they compensate each other, and so there is no deviation for the whole tree. To avoid such mistakes, it was decided to perform a 1: 1: 1: 2 deviation of a pendulum, starting from the smallest one, which contains the genes of all four species. To accomplish this task, a script was written that exploited the ability of ETE to read the structure of the tree and a particular search template was set. This was accomplished in the following way:

First, the objects of interest are defined with the names of the species and the first few letters of the genes that define their species. Then, the expected 1: 1: 1: 2 ratio is defined. The search for deviations from this ratio should start from the leaves and continue towards the root.

If a gene (leaf) is from one of the four tested species, the smallest tree (subtree) that includes genes of the other three species is selected, and the ratio of the genes in this subtree is calculated. Then, if the tree matches the standard ratio, the subtree is ignored and the search continues downwards. If a deviation is detected, the genes that cause this deviation are recorded with their corresponding rows in the informative part of the tree. Then the steps are repeated until the root of the tree is reached, and the DATA row, the structural part (the Newick tree), and the separator (//).

Due to the nature of the script reading a tree from the leaves to the root, some genes are present into the list more than once due to the consideration of the cloud, which includes the previous one. This requires writing an extra script to crawl the resulting file that searches for and removes the reps, leaving only unique genes.

3 Results and discussion

After completing the theoretical preparation of massive trees, those with at least one gene of the four species studied were extracted. From these trees, the script designed to look for deviations in the standard quantity between genes produced a list of discovered duplicated genes. After removing repetitive elements, 4322 duplicated genes were found, which are found in 1493 trees (figure 1). The distribution of genes by species is shown in Figure 2A.

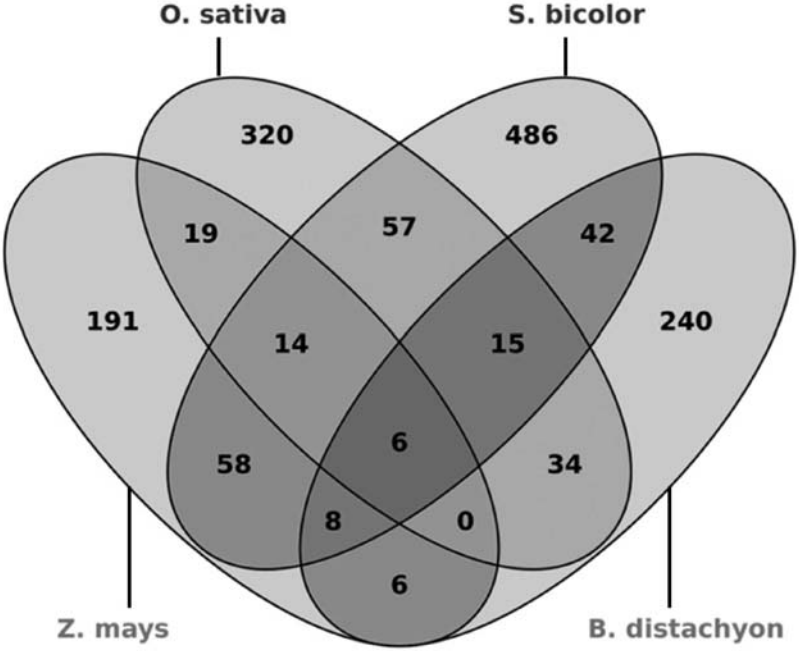


Figure 1: Distribution of trees containing duplicated genes, depending on the number of species to which the deviation is attributed.

For a clearer visualization of the results, the list was displayed in comma-separated values, which can be read from both a text editor and Microsoft Excel. It is a list of informational rows for genes that do not match the given relationship, followed by the tree identifier. Inserting the identifier saves Newick's addition of the tree itself, which would only unnecessarily load the source file. In order to facilitate the analysis, the information fields were formatted in separate columns as the information is used in the next steps.

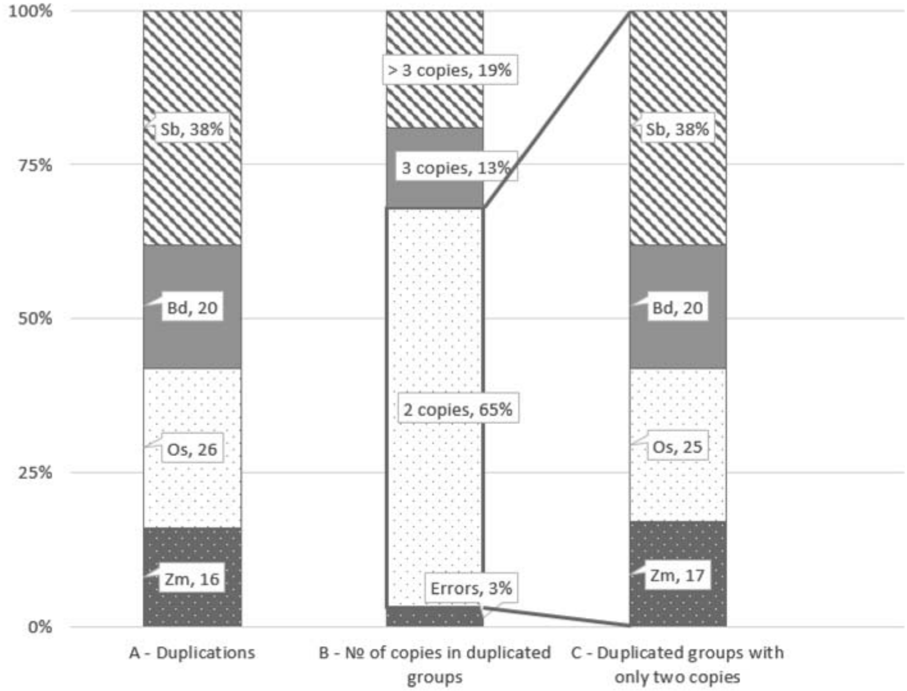


Figure 2: Distribution of duplicated genes leading to pattern deviation. A shows distribution of all diplicated genes by species; B – the size (number of genes) of the groups, and C represents the distribution of genes only in the groups with two copies.

As seen in Figure 1, most deviations from the standard ratio come from the duplicated genes of one species (82%). The other deviations occur in two or more species. In order to proceed to the next stage of the analysis, all these trees first have to distinguish between the pairs and the larger groups of duplicated genes.

When only two or three genes belonging to a species are present within one tree, they can be unambiguously referred to the same group. When the genes are four or more, it becomes necessary to specify whether they are a single group of duplicated genes or should be considered as separate groups. This is done by assuring that each set of genes forming a group are located on the same chromosome and read in the same direction. To do this, the following validation is done: 1) location of the genes in the genome - the duplicated genes are located on the same chromosome, which can be checked in the Chromosome field of the informative part; and 2) the direction of reading the DNA strand - in order for subsequent analyzes to be performed properly, it is necessary that the entire group of duplicated sequences be oriented in the same direction. This is checked in the Strand field of the informative part.

Once all duplicated group are defined, the analysis can continue by comparing the expression levels of the genes in each group within leaf tissues.

4 Conclusion

The current *in silico* approach addressing the evolution of C4 traits relies on finding and tracing a repeatable pattern in the topology of trees containing genes form well annotated C3 and C4 cereals. The results are going to be validated by comparing the expression levels of duplicated gene groups – an approach used by other authors in the same field. An additional validation could be carried out by comparing the topology of predicted candidates with that of referent genes whose role in C4 photosynthesis is experimentally confirmed.

This evolutionary approach is an alternative to most other studies on C4 photosynthesis that rely on sequence analyses of a limited number of genes and genomes. The study is entirely based on public datasets which saves both time and resources, and discovers new knowledge in the results of different experiments.

The authors' method for pattern discovery in the topology of phylogenetic trees can be easily modified to address other alternating phenotypes.

5 Acknowledgements:

The presented work has been funded by the Bulgarian NSF within the "Methods for Data Analysis and Knowledge Discovery in Big Sequencing Datasets" project, Contract DFNI-I02/7 of 12 December 2014.

6 References

[1] Sage, R. F. (2004). The evolution of C4 photosynthesis. *New phytologist*, *161*(2), 341-370.

[2] Kersey, P. J., Allen, J., Christensen, M., Davis, P., Falin, L. J., Grabmueller, C., Seth, D., Hughes, T., Humphrey, J., Kerhornou, A., Khobova, J., Langridge, N., McDowall, M., Maheswari, U., Maslen, G., Nuhn, M., Ong, C. K., Paulini, M., Pedro, H., Toneva, I., Tuli, M. A., Walts, B., Williams, G., Wilson, D., Youens-Clark, K., Monaco, M. K., Stein, J., Wei, X., Ware, D., Bolser, D. M., Howe, K. L., Kulesha, E., Lawson, D., Staines, D. M. (2014). Ensembl Genomes 2013: scaling up access to genome-wide data. Nucleic acids research, 42 (D1): D546-D552.

[3] Food and Agriculture Organization of the United Nations. (2014). FAOSTAT statistics database.:FAO (http://faostat3.fao.org/home/)

[4] Brkljacic, J., Grotewold, E., Scholl, R., Mockler, T., Garvin, D. F., Vain, P., & Caicedo, A. L. (2011). Brachypodium as a model for the grasses: today and the future. *Plant Physiology*, *157*(1), 3-13.

[5] Swigoňová, Z., Lai, J., Ma, J., Ramakrishna, W., Llaca, V., Bennetzen, J. L., & Messing, J. (2004). Close split of sorghum and maize genome progenitors. *Genome research*, *14*(10a), 1916-1923

Model of Information System for Company Management According to Software Engineering Standards ISO/IEC 12207:2013

Katerina Dimeska[1], Snezana Savoska[2],

[1]Pro grupa – Bitola, [2]Faculty of Information and Communication Technologies
University „St.Kliment Ohridski“ – Bitola,7000,ul. Partizanska bb, Macedonia,
katerina@progrupa.mk, snezana.savoska@fikt.edu.mk,

Abstract. The standards that govern the quality of life cycle processes of the system/software are led by the ISO/EIC (IEEE) 12207 standard (System and Software Engineering: 2008), which establishes a framework that contains processes, activities and tasks that need to be applied in software product and service acquisition and software supply, development, operation, maintenance and disposal. The base version of standard is regularly updated and known as ISO/IEC 12207: 2008. With the recent changes, it can be found as AS/NZS ISO/IEC 12207: 2013 and includes 17 processes, of which 5 are primary life cycle processes of the system, 8 support processes and 4 organizational processes. The standard main purpose is providing a common structure, so buyers, suppliers, developers, maintainers, operators, managers, and technicians involved in software development use a common language. This common language is established as form of well-defined processes. The standard structure has flexible design, providing modular way of tailoring the needs of users. The standard is based on two basic principles: modularity and responsibility. Modularity means that the processes have minimum joining and maximum cohesion. Responsibility means establishment of responsibility for each process, facilitating application of projects' standard with many people involved. The paper propose a model of wide range information systems for company management development, according to the mentioned standard and their principles.

Keywords: ISO/EIC (IEEE) 12207, Standards for software engineering, System development life cycle processes.

1. Introduction

There is an essential need for software systems to satisfy today's complex and costly business demands. The standard is a document with technical specifications or other criteria in order to ensure that the material or method will consistently meet the intended needs.

The goals of standardization in information technologies are: defining a common framework that will enable everyone involved in the process of development, design and management of the software to "speak the same language", providing a basis for communication between information systems (ISs), preconditions for joint different parties participation in projects and framework for software with

V. Dimitrov, V. Georgiev (Editors): ISGT'2017. ISSN 1314-4855
Proceedings of the 11th International Conference on
Information Systems and Grid Technologies, Sofia, 29.-30. Sep. 2017.

defined quality development as well as its implementation. The main principles of standardization in information technology are:

- Standard does not prescribe a specific model of life cycle of software or methods development
- Standards - applying parties are responsible for choosing a life cycle model and mapping processes, activities and tasks into selected model
- Parties are also responsible for selection and application of software development methods and for execution of activities and tasks corresponding to the software project.

Software process standards (as ISO/IEC 12207) and models (as CMMI) have been developed by international associations for helping to software development organizations to meet the current demands for quality process and software product improvements.

The ISO/IEC 12207:2013 standard is one of the fundamental standards of software engineering. It describes the architecture of the life cycle of software from concept to withdrawal. This standard is applied to software products and services in the procurement, delivery, development, use and maintenance and it represents the first internationally accepted framework for software projects implementation.

This standard is voluntary that means, it does not impose any obligation upon anyone to follow it. Yet, it may be imposed by an organization through internal policy directive or by individual parties through contractual agreements. The standard is designed for use by one or more parties as the basis of an agreement or in a self-imposed way.

An organization that is not using this ISO standard, is likely to have a chaotic way of producing software and dissatisfied customers.

The primary goal of the paper is to create a model of information system for company management according to software engineering standards ISO/IEC 12207:2013. The paper is organized as follows. Section 2 take into consideration the mentioned standard. Section 3 describe the model development processes according to the Standard recommendations. Next section explain software implementation processes. Finally, the explanation of the proposed model according to ISO 12207 standard is highlighted in section six, followed by concluding remark of the paper.

2. The ISO/IEC 12207 Standard

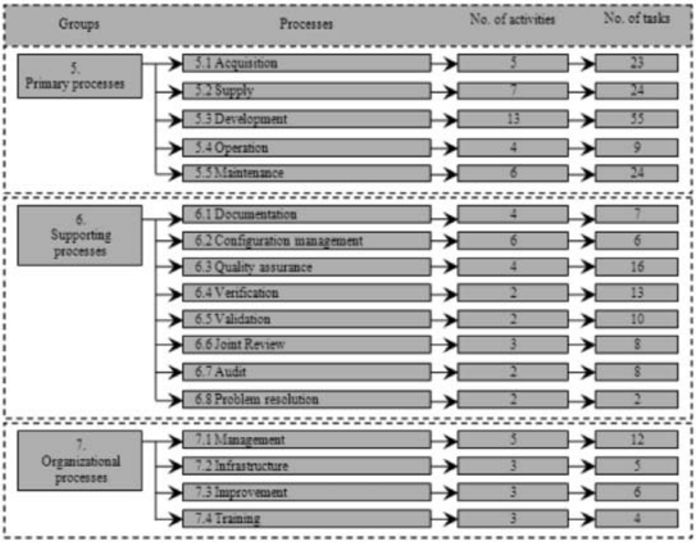
The ISO/IEC standard 12207 identifies mandatory processes, tasks and activities for software life cycles. It is intended to be applicable to software development in a broad range of application domains and for a variety of different software systems. It contains a normative and an informative component. The normative component of ISO/IEC 12207 determines mandatory practices that have to be followed by a particular development effort, compliant with Standard. The informative component identifies rationales for the practices required by the standard and explains their application.

ISO/IEC 12207 define the software life cycle processes (SLPs) classification. This standard is flexible, modular and compatible with whole software life cycles [12].

This standard is defined at the level of process rather than procedure. Rather than provide the step-by-step requirements characteristic of a procedure, it describes continuing responsibilities that must be achieved and maintained during the life of the process [11].

ISO/IEC 12207 covers the entire life cycle from "conceptualization of ideas through retirement". It considers the software life cycle from different levels of abstraction. At the highest level, it identifies a number of processes as shown in Figure 1. Three different categories of processes are identified [3]:

1. *primary life cycle processes* are conducted by prime parties, i.e. those that initiate or perform the development, operation or maintenance,
2. *supporting life cycle processes*, for instance configuration management; these support primary processes in order to contribute to the success and quality of the project, and
3. *organizational life cycle processes* that are employed by an organization to establish and implement the underlying structure of the life cycle, such as management and process improvement.



Groups	Processes	No. of activities	No. of tasks
5. Primary processes	5.1 Acquisition	5	23
	5.2 Supply	7	24
	5.3 Development	13	55
	5.4 Operation	4	9
	5.5 Maintenance	6	24
6. Supporting processes	6.1 Documentation	4	7
	6.2 Configuration management	6	6
	6.3 Quality assurance	4	16
	6.4 Verification	2	13
	6.5 Validation	2	10
	6.6 Joint Review	3	8
	6.7 Audit	2	8
	6.8 Problem resolution	2	2
7. Organizational processes	7.1 Management	5	12
	7.2 Infrastructure	3	5
	7.3 Improvement	3	6
	7.4 Training	3	4

Figure 1. Structure of the International Standard ISO 12207 [3]

Processes are decomposed into activities. The acquisition process, for instance, is decomposed into activities for initiation, request for tender, contract preparation, supplier monitoring, and acceptance and completion. Activities are further decomposed into tasks. The request for tender preparation activity, for instance, encompasses tasks determining system requirements, the scope of the system, instructions for bidders, general terms and conditions, control of subcontracts and technical constrains [9].

As the standard is meant to be applicable in many different domains, it covers a variety of processes, activities and tasks [8]. In order to define customizations of the standard for a particular domain, organization or project, the normative part of the standard includes a tailoring process. It defines how the standard have to be adapted and indicates those processes, activities and tasks that might be omitted. Moreover, it allows processes, activities and tasks to be added, provided that they are specified in a way that is compliant with the standard. The ability to tailor the standard allows its application in a number of different settings, such as waterfall, evolutionary, incremental and spiral process models.

2.1. Principles of the standard

The standard is based on two basic principles: modularity and responsibility. Modularity means processes with minimum coupling and maximum cohesion. Responsibility means to establish a responsibility for each process, facilitating the application of the standard in projects where many people can be legally involved.

2.2. Standard's basic requirements

This International Standard contains requirements in four Clauses: Clause 6, which defines the requirements for the system life cycle processes, Clause 7, which defines the requirements for specific software life cycle processes, clauses of Annex A, which provides requirements for tailoring of this International Standard and clauses of Annex B, which provides a Process Reference Model (PRM) which may be used for assessment purposes [4], [10].

2.3. Standard's Limitations

This International Standard describes the architecture of the software life cycle processes but does not specify the details of how to implement or perform the activities and tasks included in the processes [2]. It is not intended to prescribe the name, format, or explicit content of the documentation to be produced. Also, it may require development of documents of similar class or type, various plans are an example. International Standard, however, does not imply that such documents have to be developed or packaged separately or combined in some fashion. These decisions have to be made by the user.

Also, this Standard does not prescribe a specific life cycle model or software development method. The parties of this International Standard are responsible for selecting a life cycle model for the software project and mapping the processes, activities, and tasks in this International Standard onto that model. The parties are also responsible for selecting and applying the software development methods and performing activities and tasks suitable for the software project.

This International Standard is not intended to be in conflict with any organization's policies, standards or procedures that are already in place. However, any conflict needs to be resolved and any overriding conditions and situations need to be cited in writing as exceptions to the application of this International Standard.

2.4. Standard Purpose

The purpose of the Standard is to provide a defined set of processes to facilitate communication among acquirers, suppliers and other stakeholders in the life cycle of a software product. It is written for acquirers of systems and software products and services and for suppliers, developers, operators, maintainers, managers, quality assurance managers, and users of software products. Its limitations does not detail the life cycle processes in terms of methods or procedures required to meet the requirements and outcomes of a process. Its objective is to provide the software industry with a common framework for software life cycle processes and provides a process that can be employed for defining, controlling, and improving software life cycle processes [2].

3. The model development processes

ISO/IEC 12207 standard can be used in one or more of the following modes [1]:

- By an organization - to help establish an environment of desired processes. These processes can be supported by methods, procedures, techniques, tools and trained personnel. Organization have to create a suitable model for software development and employ the environment to perform and manage its projects and progress systems through their life cycle stages. In this model, the Standard is used to assess conformance of a declared, established set of life cycle processes to its provisions.
- By a project - to help to select, structure and employ the elements of an established set of life cycle processes in order to provide products and services. The created model have to support these activities. In this model, the Standard is used in assessment of conformance of the project

to the declared and established environment.

- By an acquirer and a supplier - to help develop an agreement concerning processes and activities. Through the agreement, the processes and activities in the Standard are selected, negotiated, agreed and performed and then suitable model have to be set. In this model, the Standard is used for guidance in developing the agreement.
- By organizations and assessors - to perform assessments that can be used to support organizational process improvement and these activities have to be the base for the created model.

The software process standards and models are important for software development organizations because their implementation has generated benefits such as: cost reduction, fit to schedule, better quality and better customer satisfaction [5]. Furthermore, modern business international practices demands from business organizations (including software development organizations) to provide evidences on the quality, consistency and standardization of their used processes for delivering their products and services. Thus, business organizations (including software development ones) are highly interested in implementing and using a software process model or standard.

The processes in the standard form are a comprehensive set intended to serve various organizations. The organization, small or large one, depending on its business purpose, can select an appropriate subset of the processes (and associated activities and tasks) to fulfill that purpose. Organization can perform one or more processes. A process have to be performed by one or more organizations. The standard is intended to be applied by an organization internally or contractually by two or more organizations. In order to facilitate application of the standard either internally or contractually, the tasks are expressed in contractual language. When the standard is applied internally, the contractual language is interpreted as a self-imposed task.

The organization have to manage its processes according to management processes plan, establishing the infrastructure under the processes, providing staff training and improving the corresponding processes. Of particular importance is process improvement because once the organization established the infrastructure of the needed processes and staff training, it performs continuous improvement based on the applications of these processes [5].

Improving software development or software processes is a primary challenge for present community since the software has become the core and the crucial component of any modern service or product. Therefore, ensuring its quality is essential and should not be ignored. Consequently, software process improvement (SPI) is one of the most important and critical efforts that any software development organization pursues [6].

4. Software implementation process

If not stipulated in the contract, the developer shall define or select a life cycle model appropriate to the scope, magnitude, and complexity of the project. The implementer shall document the outputs in accordance with the software documentation management process and place the outputs under the software configuration management process as well as to perform change control management. After that, he shall document and resolve problems and non-conformances found in the software products and tasks in accordance with the software problem resolution process and perform supporting processes as specified in the contract. Establish baselines and incorporate configuration items at appropriate times, as determined by the acquirer and the supplier [7].

Couple of various software development process models can be selected (e.g. Waterfall model, incremental model, V-model, iterative model, etc.). The methodology within the SDLC process can vary across industries and organizations, but standards such as ISO/IEC 12207 represent processes that establish a life cycle for software, and provide a model for the development, acquisition, and configuration of software systems. In this paper we will take a look at the V- model which follows a particular life cycle in order to ensure success in process of software development which is shown in Figure 2.

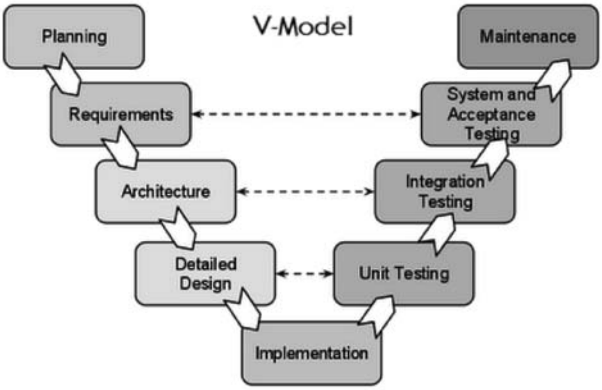


Figure 2. V – Model of ISO 12207 [13]

5. Explanation of the proposed model

The V - model is SDLC model where execution of processes happens in a sequential manner in V-shape. It is also known as Verification and Validation model. V – Model is an extension of the waterfall model and is based on association of a testing phase for each corresponding development stage. This means that for every single phase in the development cycle, there is a directly associated testing phase. This is a highly disciplined model. The model is sequential, that means the next phase can start only after the completion of previous phase. Such models are suitable for projects with very clear product requirements and where the requirements will not change dynamically during the course of project completion.

The advantages of this model include: It is ease of use, testing happens at each phase of the process, increasing the chances of success as any defects can be identified and fixed before coding. Also, the model works well for smaller projects. There is some disadvantages of this model that include: Inadequate for large and complex projects, inflexible model, just as waterfall model, because no early version of model were produced, so the bugs can be discovered after the model is developed.

Taking into consideration the possible improvement of the model with usage of Standard ISO/IEC 12207:2013, the proposed activities includes incremental improvement of procedures, process establishment, process assessment as well as process improvement processes (Figure 3).

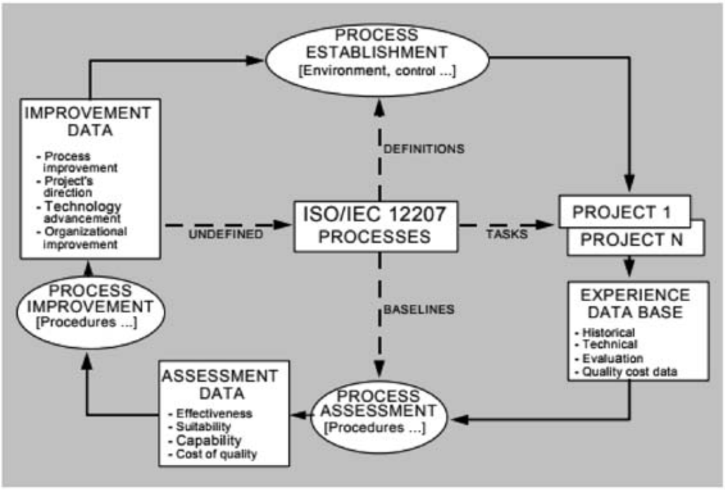


Figure 3. Process improvement processes with ISO 12207 [8]

6. Conclusion

The use of standards for software engineering ISO/IEC 12207:2013 has many potential benefits for any organization, such as improved software management, improved certification visibility that can attract a new customers and partners, enhancing partnerships and co-development etc., particularly in a global environment.

This standard is very important for companies' information system management with encapsulation of best practice, avoiding some common past mistakes, establishing a framework for quality assurance process and providing continuity.

ISO/IEC 12207 is the first International Standard that provides a complete set of processes for acquiring and supplying software products and services. These processes may be employed also to manage, engineer, use, and improve software throughout its lifecycle. Its architecture can accommodate evolving, modern software methods, techniques, tools, and engineering environments. The expectation is that ISO/IEC 12207 would fulfill its intended purpose as the basis for World trade software products and services.

7. **References**

[1] ISO/IEC JTC 1/SC 7/WG 7, ISO/IEC 12207:2008(en), Systems and software engineering — Software life cycle processes, 2008

[2] AS/NZS ISO/IEC, Information technology - Software life cycle processes, 2003

[3] Al-Qutaish, R.E. Measuring the Software Product Quality during the Software Development Life- Cycle: An International Organization for Standardization Standards Perspective, Journal of Computer Science, 2009

[4] Lewis Gray, Guidebook to IEEE/EIA 12207: Standard for Information Technology, Software Life Cycle Processes, 2000

[5] Michael E.C. Schmidt, Implementing the IEEE Software Engineering Standards, 2000

[6] Shamsul Sahibuddin, Software Process Improvement and Management, 2011

[7] A. Riel, Systems, Software and Services Process Improvement, 2010

[8] Raghu Singh, An introduction to ISO/IEC 12207 (Tutorial), 1999

[9] Alan Jones, ISO 12207 Software life cycle processes – fit for purpose?, Software Quality Journal 5, 1996

[10] Standards Australia Limited/Standards New Zealand, Systems and software engineering - Software life cycle processes, 2013

[11] James W. Moore, IEEE/EIA 12207 as the foundation for enterprise software processes, 2000

[12] Samira Haghighatfar, Presentation of an approach for adapting software production process based ISO/IEC 12207 to ITIL Service, 2013

[13] Mohamed Sami, Software Development Life Cycle Models and Methodologies, 2012

Big Data Visualization Concepts and Visual Data Analytics

Snezana Savoska[1], Dragan Milevski[1]

[1]Faculty of Information and Communication Technologies
University „St.Kliment Ohridski" – Bitola, 7000, R. Of Macedonia,
snezana.savoska@fikt.edu.mk, dragan_055@hotmail.com,

Abstract. The latest research made by the World business researchers, shows that the Visual Data Analysis tools market is constantly growing and includes more software companies. It is important when the analysts are taking into consideration the competencies and ease of use, the compatibility with the tools for data collecting, storing and processing, information presentation as well as gained knowledge from data. The emerging data trends are connected with Big data concepts and the manner of their preparation for analysis and creating visual data analysis. The purpose of this paper is to consider concepts of Big data visualization analysis and supporting software tools. Considering Big data characteristics, general "big picture" visualization can be gained. The visualization depends on many factors. Some standards have been established in the context and show how the data has to be prepared for visualization, using integrated data visualization tools and Business Intelligence.

Keywords: Big Data Visualization, Visual Data Analysis, Information visualization, Zeppelin.

1. Introduction

The rapid data flood leads to a data rush stored in the companies' database, on servers or on cloud [15], [19]. The data collected in this way, stored in the different formats and platforms, various databases, distributed or stored in companies' data warehouse are not useful for data analysis and cannot be used for extracting useful information about trends, insights and exceptions. They do not provide some visible information and knowledge that can be used as a tool for gaining a competitive advantage and getting new customers. Confronted with this serious problem, the biggest world data science companies produce software for dialing with this huge data flow. First, they introduced the term "Big Data" used to show data with 6V characteristics [2], [14] and starting to explore in the Data science frames, tending to use all capabilities of Data analysis science using emerging technology of artificial intelligence, mathematics, statistics methods as well as data mining in databases. Some researchers start to deal with this data flow using visualization methods (Visual Data Analytics - VDA), one of the most desired methods used by business community because of the rapid visual human eye information processing [4]. VDA techniques aim to visually analyze data using software and methodologies for preparing data in the most appropriate way to achieve the maximal level of data representation according to Shneiderman's principles [19]. VDA tools market's demands increase because of inclusion of Big data in visual analysis and Business Intelligence (BI) as well as the need of many kinds of heterogeneous data analysis in short time. The market demands for VDA software lead to the fact that this software industry includes more and more business players using power capacities for improving their software tools for covering Big Data Visualization concepts with VDA tools. The trends connected with data, included in Data science, lead to the concept of Big data and the way of preparing data for analysis, creating VDA, using methods of visualization and supporting tools.

In order to answer the question if Big data can be directly visualized and how the Shneiderman's concept can be used for getting information from data [19], first overview, second zoom and select, then detail on demand, the question that arise is how and with which tools the general "Big picture" visualizations has to be gained and then, how some interaction with data has to be used for getting detail on user's demand.

The next statement that has to be explained is that the visualization process depends a lot on the data type, the storage place, data scalability and structure as well as the display size, computer's processing power, the data scientist knowledge, communication, cooperation and shared user's needs, that all influence on the way of data preparing for visualization and effective and efficient representation on the users' screens [5]. Some of the important factors are taken into consideration in the paper.

The paper is organized as follows: After the introduction session, overviews of the recent research from this area is provided and some "pictures" of the recent research in Big data visualization and Visual data analytics are taken into consideration. The next two sections provide some reviews of used concepts and methodologies for VDA with the indicators, needed for objective assessment of software solutions for VDA of Big data. The conclusion remark summarizes the researches in the area and give a trends' overview of VDA and Big data Visualization.

2. Related works

Many research is committed on the use of visualization for data analysis [4], [19], [9]. In the resent years, the focus is on Data Science that includes usage of techniques and methodologies for dialing with data, using a wide range of Big data concepts and algorithms for data analysis, software for Big data analysis and visualization [2], [23]. Focus on Data science provides fast movement and development of tools that provide big power for decision makers, empowered with right and timely information [10]. This huge data flow analysis provide trends identification and future events prediction in order to add value to the organization, using Business Intelligence and Analytics (BI&A) systems that transform row data into valuable information in the visual manner, named Visual Business Intelligence (VBI) [21]. According to these research, these tools provide more sense in data, data sets are presented in visual form, some hypothesis are tested, some trends are emphasized and insight in data are made [2].

All amenities that VDA usage provide, increase the demands for qualified staff, capable for working with data and VDA, especially in USA and the developed countries, that means existing of deficit of personnel with specific skills to work with data and VDA tools [8]. This also means increased demands for managers who have knowledge for making decisions based on evidence obtained with VDA of Big data [10].

Chen [8] states that VDA are techniques, technologies, methodologies, tools and practices that analyze critical business data in order to help for better understanding of their businesses and markets and providing timely business decisions. VDA is emerging area, developed on the base of statistic, mathematics, probability and data representation in order to get a sense from big data sets, stored in databases [21]. Many authors emphasize that the companies today depend more and more on decision support tools, BI&A and visualization [3]. Turban [20] and Chaudhuri [7] state that, with these concepts, many business processes are improved in many big companies. Chen [8] connects BI&A with Big data Analytics that become increasingly important for academic as well as business community in recent decade. All these facts cause staff demands with enhanced analytical skills [17] and many universities introduce courses to help in VDA and Data visualization, based on the concepts of learning-by-doing and trial-and-error experiments [8].

Aigner&all. [1] research how much the interactive visual methods are used in Austria. From quantitative research led through semi structured interviews with VDA users and usage of cognitive and post cognitive methods and theories, they conclude that usage of visual analysis mostly depends on corporate (organizational) culture, job attractiveness and creativity. He posed hypothesis that VDA use statistical analysis with limited interaction and that the users looking for more interaction because interactivity helps in gaining more information and knowledge from data with VDA, especially when VDA on Big data is taken in consideration [1]. However, moving toward Big data concept inevitably requires others visual analysis types of tools that outweigh current used techniques and demand interactivity in order to deal with data complexity and possibility to compare many scenarios and alternatives and deeper results understanding needed in the decision making processes.

3. **Concepts and Methodologies Used for VDA on Big data**

The symbiosis of VDA and Big data brings a new insights and observation in huge amount of data, collected from different types of interactions in business and human interactions in general. The purpose is to obtain information and knowledge that have to be used in gaining competitive advantage and bigger market share [11]. Companies see the information as corporate wealth and the visualization is used as visual point-and-click querying, dashboard development tools, intelligent natural languages similar to Google interface or self-services tools [18]. Analysts are studying the tools of Data science, VDA, BI&A according to market, technology and allowed conditions as well as according to execution, market fit and marketing mix (Fig.1). The analysis gained from the research is different, the different visualization techniques are used because some other analytical methods are used with techniques as stacked bar, radar char, calendar hit map, chord diagrams, theme river visualizations, tree maps view, social network and many other visualization types. Ovum analyzes the integration of Big data concept with data visualization and points out haw Hadoop and NoSQL databases tools have to be used to create a new frame for the next database generations (Fig.2), the analysis previously impossible to create because of huge data volume and non-supported performances in this concept [18].

IBM uses Big data visualization to obtain the general picture of this data [16], because the human visual system could process 9 MB information per second that corresponds to 1 million letters text per second. They are taking into consideration data collected with RFID communications, social media, customers' data reviews stored in long time period, Internet of things (IoT) data and others. According to their research, only 26% of companies can analyze unstructured data as sound and vision and only 35% are capable to analyze streaming data. They argue that the most important is to gain a "Big picture of data" which cannot be created with ordinary Dashboard tools and reports tools using Key performance indicators (KPI) and historical data.

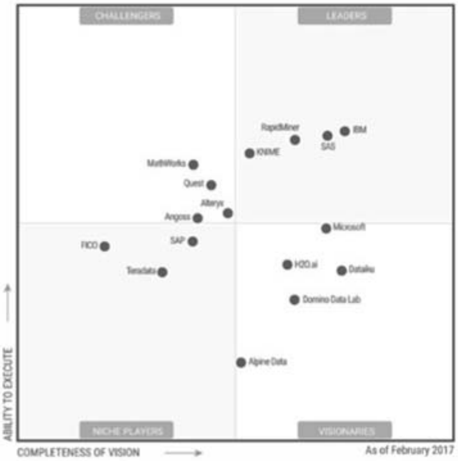


Fig. 1. Gartner 2017 magic quadrant for Data Science platform [13]

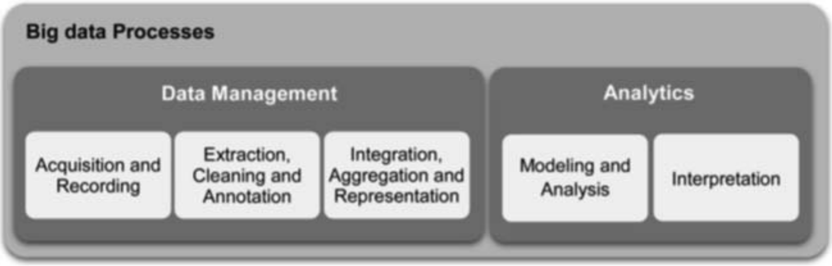

Big data Processes
Data Management
Analytics
Acquisition and Recording
Extraction, Cleaning and Annotation
Integration, Aggregation and Representation
Modeling and Analysis
Interpretation

Fig. 2. The data extraction and data preparation for visualization [12]

With increasing columns number database and number of data, it is obvious that the usual tools cannot meet the challenges. Sometime, data reduction with segmentation, clustering, linear regression or logarithmic methods tools are used before creation of visualization in order to cover the essential data characteristics from 6V (Volume, Variety, Velocity, Veracity, Variability, and Value) [14]. Adding the time, some other data types have to be created, suitable for time series analysis or sessional analysis with Radar charts and Calendar heat maps. On the other hand, in order to understand the customers' feelings, many other Big data are used, as the call centers' logs, social networks (Fig.3), social media analysis (Fig.5) and customers surveys (Fig4). Some methods as extract patterns have to be used with techniques as Theme River Visualization or tree maps (Fig.5), which can be used to detect relationships between customers. Other structures of hierarchical splitting or tree structures can display millions nodes and relations between them (Fig.3).

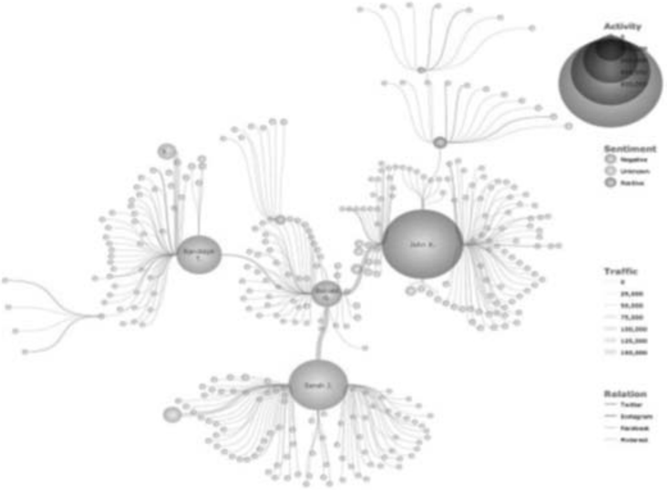


Fig. 3. Social networks visualization, customers' fillings patterns, key influences and their bands [16]

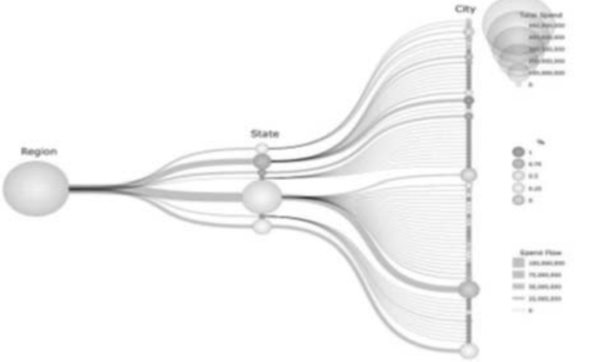


Fig. 4. Hierarchical data visualization showing the answer number of target companies on regional, state and city's level [16]

4. Strategies and Tools for VDA of Big data

One highlighted question is "Can Big data be directly visualized in database?" and this question causes many controversies about what can be achieved by this "Big picture" visualization. Certainly this visualization depends on data scalability and data structure as well as display size, processor power and the user's needs [15] [19]. Some standards show 12 iterations for terabyte and 15 iterations for petabyte Big data when preparing data for visualization and this causes integration of DVA tools with Business Analytics Solutions (BAS) [20]. One of the VDA and BAS solutions, IBM created Rapidly Adaptive Visualization Engine (RAVE) which includes new types of diagrams, unknown before and under the development phase. They products use RAVE libraries to enable interactive data visualization and solutions which simplify Big data visualization itself.

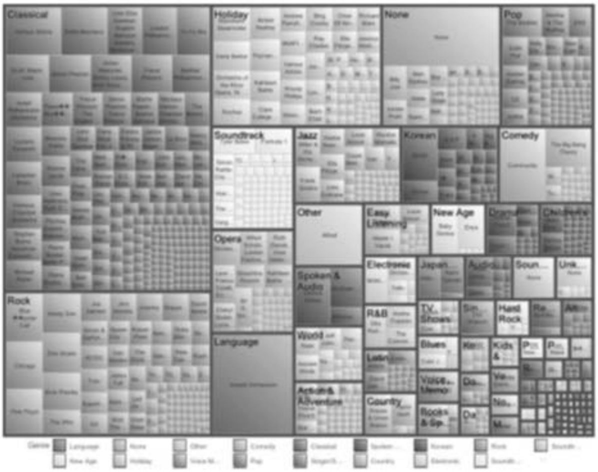


Fig. 5. Tree map visualization for selection of type of music as streaming data in social networks [18]

The global companies admit that they had a huge problems with their data understanding as well as customers' and vendors' data understanding and they got annoyed as they have established a new form of reporting, associated with a lot of effort and time spent [22]. Big data analysis was achieved with creating a framework that includes Advanced Data Visualization tools with previous defined semantic definitions, data modeling and Data Source Mapping in order to integrate data sources that have to be visualized in the middleware and to bring Actionable intelligence for customers [6]. Usually they implement Apache Hadoop, eXtrime Data Warehouse and Big Data Analytics with creating a focus of self-service BI tools, supporting mobile devices and rich and consistent experience forms [22].

However, today's methodologies and techniques are connected with VDA of Big data, are more related with usage of Hadoop, MapReduce and Spark as the next generation of paradigm for processing of Big Data, helped with visual tools for models development, using artificial intelligence with machine learning, integrated with its MLlib library. Although Apache Spark integrate some Java, Scala and Python opportunities, its graphic tools as Seahorse are used to enable enhanced visual data analysis. The results obtained with data preparation can be modeled and some iterative preparation can be shown in order to create better models for predicting some situations using Machine learning models with training algorithms based on previously collected data [21].

The other paradigm for Big Data analysis and visualization is Apache Zeppelin as open source web based tool that enable interactive data analysis and visualization. It is multifunctional software tool enabling data loading, data exploring, data visualization and data collaboration, also using Hadoop и Spark. It is a power tool for engineers and Data scientists enabling more productive results as they create some outcomes in the development phase, because of easy data exchange, analysis and visualization. It is interactive tool enabling long lasting workflow processes which can be connected with Spark, supported by Python codes as well as Scala, Hive, SparkSQL, shell and markdown [21],[22].

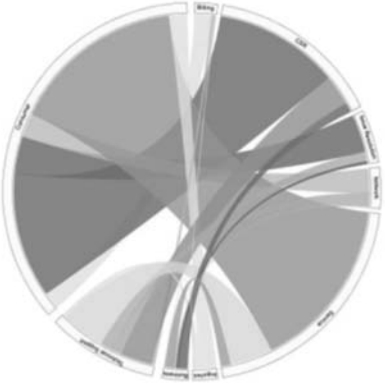


Fig. 6. Chord Chart created with IBM RAVE advanced tools for Big data visualization [16]

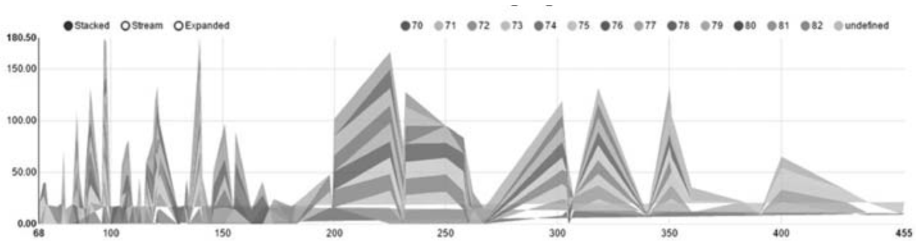


Fig. 7. Overview Areas graph for cars power related to minimal acceleration, grouped by year of production

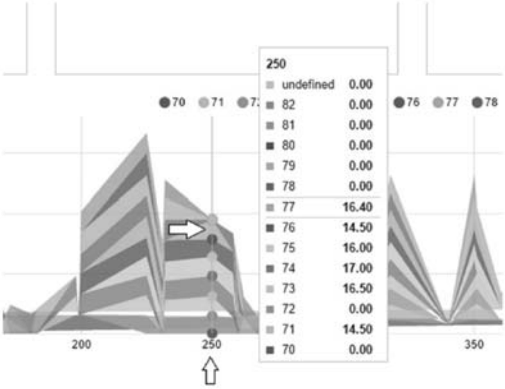


Fig. 8. Possibility of interactive Detail on demand of cars from database cars.csv

For this purpose, some Big data visualization with Zeppelin are made in order to gain some data insights. Data are not previous prepared, just they are taken from database in csv format and some interactive visualization are made (Fig.7). Interaction with data is enabled using available data features for cars. csv (Fig.8) or filter (Fig.9). VDA enable first overview, selection of data and interaction, according to Shneiderman's [4] visual mantra.

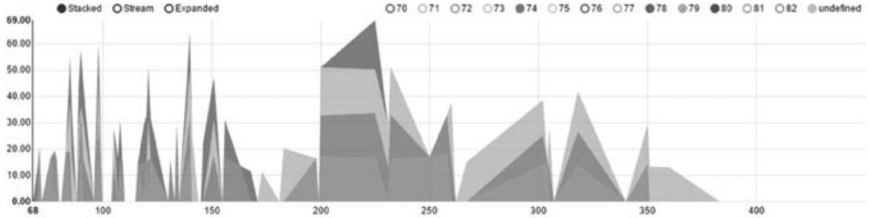


Fig. 9 Filtering on demand – years of production of cars which have to be analyzed

5. Conclusion

The Big data analysis using visual techniques and VDA include a wide range of emerging tools. Data are collected from different types of data sources, under different platforms, from many devices with automation of the processes and some parts of devices which are the part of IoT concept [22]. They are important for the organizations and companies because they believe that with their analysis, they will have information and knowledge for gaining competitive advantage and increased market, innovative ideas which lead to increasing company's profit [11]. Analyzing processes require preparing procedures with usage of different methods of Data science in order to extract useful information and knowledge from data [21]. For this purpose, a whole industry of Data science is developed in order to deal with a huge data flow in visual way [19]. The used methods for realize the concept are complex and includes sometime previous data preparation for analysis and visualization. Some tools are able to analyze data directly from database. But, many factors influence the efficiency and effectiveness of the visual representation and results gained from them [17]. The used tools are developed to collaborate and have the possibility to exchange data and some outcome [23].

6. References

[1] Aigner W., Current Work Practice and Users' Perspectives on Visualization and Interactivity in Business Intelligence, 2013 17th International Conference on Information Visualisation, IEEE, DOI: 10.1109/IV.2013.38

[2] Alfamirano A., Analyzing Big Data Using Hadoop MapReduce,

[3] Alazmi, A. R. R., & Alazmi, A. A. R. (2012). Data mining and visualization of large databases. International Journal of Computer Science and Security, 6(5), 295-314.

[4] Card, S.K., Mackinlay, J.D., and Shneiderman, B.: Readings in Information Visualization: Using Vision to Think. Morgan Kaufmann Publishers, San Francisco, pp. 1-34 (1999)

[5] Card S.K., Mackinlay J.: The structure of the Information Visualization Design Space. IEEE Xplore (2009), http://dl.acm.org/citation.cfm?id=857632, 11.01.2013

[6] Carter K.B., Actionable Intellegence, A guide to delivering business results with Big data fast, Willey, 2004

[7] Chaudhuri, S., Dayal, U., & Narasayya, V. (2011). An overview of business intelligence technology. Communications of the ACM, 54(8), 88-98.

[8] Chen, H., Chiang, R. H. L., & Storey, V. C. (2012). Business intelligence and analytics: From big data to big impact. MIS Quarterly, 36(4), 1165-1188.

[9] Chengzhi Q., Chenghu Z., Tao P.: Taxonomy of Visualization Techniques and Systems – Concerns between Users and Developers are Different. Asia GIS CP 2003, Wuhan, China (2003)

[10] Dupin-Bryant P.A., Olsen D.H., Business Intelligence, Analytics And Data Visualization: A Heat Map Project Tutorial, International Journal of Management & Information Systems – Third Quarter 2014 Volume 18, Number 3, pg.185-199

[11] Fan S., Lau R.Y.K., Zhao J.L., Demystifying Big Data Analytics for Business Intelligence Through theLens ofMarketing Mix, Elsevier, Big DataResearch2(2015) pg.28–32

[12] Gandomi A., Haider M., 2015, Beyond the hype: Big data concepts, methods, and analytics, Elsevier, Internation journal of information management, 35(2015), pp. 137-134

[13] Gartner VDA report, http://www.jenunderwood.com/2017/02/22/2017-gartner-bi-magic-quadrant-results/, пристпено, 25.6.2017

[14] Gaitanou P, Garoufallou E, Balatsoukas P. The Effectiveness of Big Data in Health Care: A Systematic Review. InMTSR 2014 Nov 7 (pp. 141-153).

[15] Gorton I., Software Architecture for Big Data Systems, Carnegie Mellon University, 2014

[16] Keahey T.A., Using visualization to understand big data, IBM Corporation, https://pdfs.semanticscholar.org/3add/2d3d3ecf7641a4cc611deecb68403ed11b33.pdf

[17] Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., & Hung Byers, A. (2011). Big data: The next frontier for innovation, competition, and productivity. McKinsey Global Institute: McKinsey and Company.

[18] Wang L., Wang G., Alexander C.A., Big Data and Visualization: Methods, Challenges and Technology Progress, DOI:10.12691/dt-1-1-7, http://pubs.sciepub.com/dt/1/1/7/, Accessed 11.8.2017

[19] SAS Institute Inc., Big Digital Data, Analytic Visualization and the Opportunity of Digital Intelligence, 2014

[20] John Wiley & Sons, Inc. Actionable Intelligence, A Guide to Delivering Business Results with Big Data Fast, 2014

[21] Apache Zeppelin: Big data prototyping and visualization in no-time, https://dataminded.be/blog/apache-zeppelin-big-data-prototyping-and-visualization-no-time, Accessed 12.6.2017

[22] Big Data and Visualization: Methods, Challenges and Technology Progress, http://pubs.sciepub.com/dt/1/1/7/, Accessed 20.5.2017

[23] Sparklet User Guide, http://mund-consulting.com/Products/Sparklet-User-Guide.pdf, Accessed 24.8.2017

BI Tools Analysis According to Business Criteria as Data Integration Possibilities, Hardware Specification, Tools for Data Visualization and Comparison of Used Technologies

Andrijana Bocevska[1], Snezana Savoska[1] , Ivan Milevski[1]

[1]Faculty of Information and Communication Technologies
University „St.Kliment Ohridski" – Bitola, 7000, ul. Partizanska bb, Macedonia,
andrijana.bocevska @fikt.edu.mk, snezana.savoska@fikt.edu.mk, ivan_055@live.com

Abstract. Business Intelligence (BI) is software tool that transforms raw data into information and knowledge, enabling managers to identify, develop or create new strategic business opportunities. Nowadays, BI tools are widely accepted, given that they provide direct access to users with intuitive information generated by real-time data, which can create a competitive advantage. This paper analyzes business intelligence software tools that are highest positioned at the Business intelligence Gartner's Magic Quadrant (QlikView, Tableau and Microsoft PowerBI) from a multi-criteria perspective in order to get an idea of the key performance indicators, such as: data integration capabilities, hardware specifications, tools for data visualization and comparison of the used technologies. The theoretical analysis is mainly based on the research from available literature on Internet and the visualizations are derived from the practical application of these software tools. The acquired knowledge in this paper will be particularly important for users interested for this kind of software to gain more knowledge when choosing the appropriate BI solution to meet their specific business needs.

Keywords: BI tools, data analysis, data integration, hardware specification, tools for data visualization.

1. Introduction

Business intelligence (BI) is the ability of an organization to collect, maintain, and organize data. Large amounts of data and information flood produced in the companies demands a proactive data analysis tools giving a new opportunities. Identifying these opportunities and implementing an effective strategy can provide a competitive market advantage and company's long-term stability [4].

BI provide historical, current and predictive views of business operations. Common functions of business intelligence technologies are reporting, online analytical processing, analytics, data mining, process mining, complex event processing, business performance management, benchmarking, text mining, predictive analytics and prescriptive analytics. [11].

The goal of modern business intelligence deployments is to support better business decision-making. Thus a BI system contains embedded part of decision support system (DSS). The companies use business intelligence as umbrella term for many similar concepts, as competitive intelligence, taking into consideration that the both concepts support decision making. BI uses technologies, processes, and applications to analyze internal and external data, structured data and business processes. Competitive intelligence gathers, analyses and disseminates information with a topical focus on company competitors. If it is understood broadly, business intelligence include the subset of competitive intelligence and many others concepts.

An extended analysis in Business Intelligence is made by Gartner and reported as the Magic Quadrant for Business Intelligence Platforms, (Fig.1) [9].

As defined by Gartner, BI platforms perform, to wide range of users, from IT staff to consultants and business users, to build applications that help organizations learn about and understand their business. Taking into consideration criteria of evaluation of BI software tools, the additional analysis is made in this paper, considering business criteria as data integration possibilities of mentioned tools, hardware specification, comparison of used technologies and tools for data visualization.

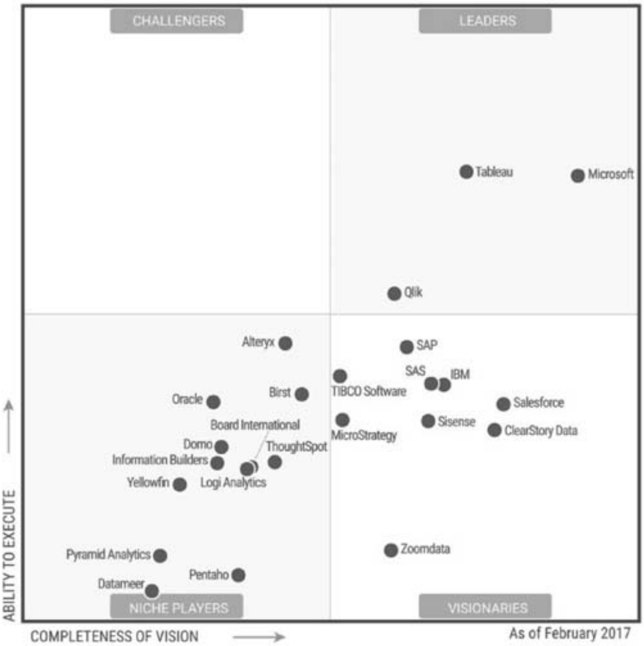


Figure 1 – Gartner's Magic Quadrant for BI platforms [9]

The paper is organized as follow. Second section takes into consideration related works and previous researches in the area. Third section makes integration overview of mentioned BI tools, while fourth section compares the needed hardware specification.

The fifth section highlights the used technology for all considered tools. The sixth section clarifies the visualization tools available for each considered software tool. Finally, the concluding section makes some final conclusion about tools' suitability for BI and their convenience for data analysis and visualization.

2. Related works

Many researches and business analysist analyze the BI market share and describe that today's Big data analysis demands powerful tools to cope with the processes of gaining information from data. Using BI tools, decision makers are empowered with many excellent tools [10, 12] equipped with additional auxiliary tools for data extracting, transforming and load as well as for visualization. These tools perform easy transformation of raw data into valuable information [13], sometime capable to represent data in visual formats, desirable for the human beings [7, 8].

Many authors state that the decision making processes demand usage of BI tools in all management companies' levels [16]. Today, many authors links BI with Big data analytics because this area is increasingly important for business community in last decade, demanding capable and educated staffs with advanced analytical knowledge [5, 15, 18]. For these reasons, leading Universities focus new curricula to Data science, data mining and other analytics tools as well as BI and Business analytic knowledge to meet business requirements [14].

Many analysts companies as Ovum see the information as company's wealth [3] and BI as strategic tools. They analyze tools according to market share, technology, assessed services, user friendliness, execution and market fit [3]. IBM analyzes tools according to possibility to get "big picture" of corporate data and then to interactive analysis of selected parameters. For them, it is important to gain a "Big Picture" of corporate data that can be done with tools as BI Dashboards, reporting tools for Key performance indicators (KPI) and Historical data. They also use some new concepts as Rapidly Adaptive Visualization Engine (RAVE) with included libraries that contain novel visualization techniques, more informing for decision makers [6].

Some researchers analyze ability and capabilities of tools for analysis and visualization of Big data in order to detect capabilities for business reports, business achievements and financial report according to KPIs, earlier defined [1, 2]. The latest attempt is made with intention to explain and highlight some KPI of Tableau and Qlik tools through comparative analysis for large number of KPI [20].

3. Integration overview

3.1 QlikView

One of the key principles of QlikView is a usage of a wide range of structured and unstructured sources of data and common associations between them (Fig.2).

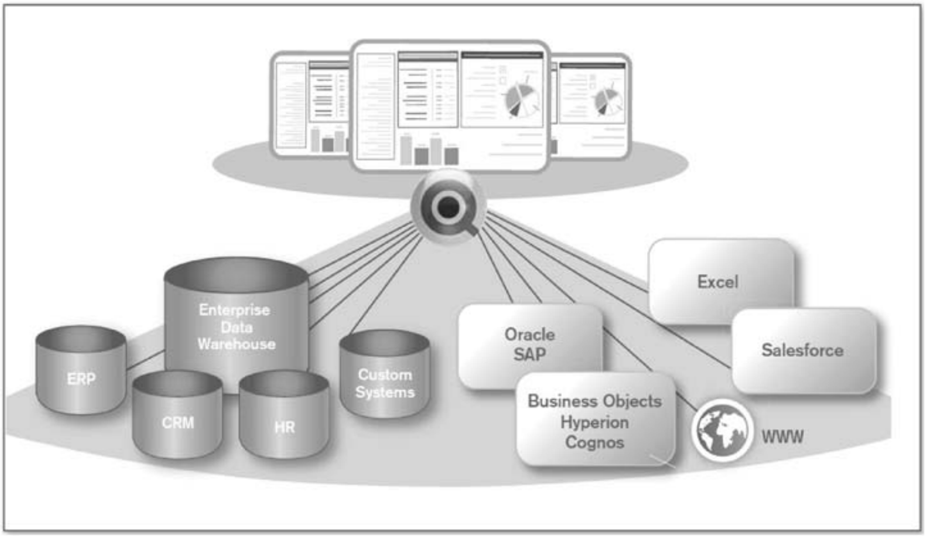


Figure 2 - QlikView can access data from a large variety of sources [17]

QlikView enables ODBC/OLE connections through a simple wizard to quickly extract data from source systems. QlikView uses standard SQL queries to read data and also use Store Procedures. It can provide access to many different types of unstructured data sources through a wide range of methods.

QlikView provides two connectors to extract data from two common data sources that do not provide typical ODBC/OLE drivers: Salesforce.com and SAP Netweaver. QlikView supports SharePoint and web parts and provide contents and document integration as objects from some portals, allowing users to get data and KPIs in dashboards along with other business content in order to provide data analysis.

QlikView also helps organizations reduce reporting costs by allowing them to rapidly develop dashboards and applications and to be flexible in dealing with business changes and demands. This include: Management integration using programmatic control of common management tasks through web service API, Automation & scheduling - Event Driven Execution (EDX) that allows a single reload to be launched and then polled for a completion status, Managing documents and source control - a single QlikView file is a binary file containing a load script, data and multiple objects in the UI, Deployment control that enables easy roll-out of multiple QlikView applications to one or more environments/ QlikView Servers using a single action.

3.2 Tableau

Tableau connects single data source with a single view for large data sets as data warehouses, data marts or flat files. The view can be different as join of multiple tables from different data source as: relational database, OLAP database, Access MDB file, Excel spreadsheet, flat file or Hadoop database in HDFS using HIVE and Apache Hadoop from Cloudera.

Users have the ability to define join operations between tables as long as they are supported by the database. If all needed data are in a single database management system, such as Oracle, SQL Server, or Teradata, database administrator can create a database view pulling data together from various schemas or user’s views.

Data can be stored in structure including transactional (3rd, 4th, or 5th normal form), de-normalized “flat” forms, and star and snowflake schemas. The Tableau view performance is directly related to the speed of the underlying structure of the database. While multidimensional databases generally perform best, a relational database with a clean star schema or an analytics-optimized database will perform higher than most other highly-normalized transaction-oriented databases.

3.3 Microsoft Power BI

One of Excel’s strongest selling points are tools (Power Pivot, Power Query, Power View, and Power Map). Characteristic of them is that after usage of one tool for analysis, the others could provide further support.

Power BI is integrated with a wide range of data sources, including both cloud and on-premises solutions. With a wide variety of data sources, it can quickly and easily connect to SaaS solutions, on-premises data in SQL Server Analysis Services, Azure services, Excel and Power BI Desktop files. Using REST APIs, you can even connect to custom data sources, such as proprietary corporate data or external data services.

One of the advantages of Power BI is that it provides access to all data from a single location, regardless of where the data resides. This hybrid approach delivers a number of benefits: (1) fast time to insight with direct connections to popular SaaS solutions; (2) secure, live connectivity to existing, on-premises data sources, such as SQL Server Analysis Services tabular; and (3) a scalable BI solution that does not require to move any on-premises data to the cloud [19].

4. Hardware specification of QlikView, Tableau and Power BI

The basic system requirements for QlikView, Tableau and Microsoft Power BI are given in Table 1. System requirements do not differ a lot. The main differences are that QlikView and Microsoft Power BI cannot run on Mac-computers. QlikView requires much more RAM memory because of its in-memory technology.

Table 1

	QlikView	**Tableau**	**Microsoft Power BI**
Platforms supported	Available for both 32 bit & 64 bit	Available for both 32 bit & 64 bit	Available for both 32 bit & 64 bit
OS Supported	Windows 7 x64 Windows 8.1 x64 Windows 10 x64 Windows Server 2008 x64 Edition Windows Server 2008 R2 Windows Server 2012 Windows Server 2012 R2 Windows Server 2016	Microsoft® Windows® 10, 8, 7, Vista, or XP sp3; or Server 2012, 2008, or 2003 (on x86 or x64 chipsets)	Windows 10 , Windows 7, Windows 8, Windows 8.1, Windows Server 2008 R2, Windows Server 2012, Windows Server 2012 R2
RAM	4 GB minimum Memory requirements are directly related to the quantity of data being analyzed.	2 GB	1.5 GB
HDD	300 MB total required to install	1.5 GB minimum	169 MB to download the software file and another 500 MB to install the software.
CPU	Intel Core 2 Duo or higher recommended	Intel Pentium 4 or AMD Opteron	1 gigahertz (GHz) or faster x86- or x64-bit

Starting with Tableau 10.5, new versions of Tableau will only run on 64-bit operating systems. Release of Tableau 10.4 is the last version of Tableau Desktop, Tableau Reader, and Tableau Public to support 32-bit Windows operating systems.

5. The Used Technologies

4.1 QlikView

The engine behind associative search is QlikView's next-generation in-memory architecture. It virtually eliminates the problems and complexity plaguing traditional, slow, disk-based and query-based BI tools that deliver more than static, prepackaged data. With QlikView, all data is loaded in memory and available for instant associative search and real-time analysis with a few clicks.

QlikView breaks out of the gridlock of the traditional BI world, where business users and developers spend months documenting and coding these requirements into dashboards, analysis, and reports, using different products for each output. Pulling data into QlikView takes just minutes because data is not required to be staged or stored in intermediary formats such as data warehouses or cubes (although QlikView can source data from these systems).

QlikView integrates both the building of the back-end underlying analytic calculations with the front-end user interface. With this complete BI solution, developers have one place to build, instead of having to use separate BI tools for dashboards, analysis and reports. QlikView application provides powerful associative search and data visualization capabilities that allow business users to view their own slice of the underlying data.

4.2 Tableau

Tableau provides two modes for interacting with data: Live connection or In-memory. Users can switch between a live and in-memory connection as they choose.

Live connection: Tableau's data connectors leverage existing data infrastructure by sending dynamic SQL or MDX statements directly to the source database rather than importing all the data. This means that if it is invested in a fast, analytics-optimized database like Vertica, the benefits of that investment by connecting live to data will be obtained. This leaves the data detail in the source system and send the aggregate results of queries to Tableau. Additionally, this means that Tableau can effectively utilize unlimited amounts of data. In fact, Tableau is the front-end analytics client to many of the largest databases in the world. Tableau has optimized each connector to take advantage of the unique characteristics of each data source.

In-memory: Tableau offers a fast, in-memory Data Engine that is optimized for analytics. You can connect to data and then, with one click, extract data to bring it in-memory in Tableau. Tableau's Data Engine fully utilizes entire system to achieve fast query response on hundreds of millions of rows of data on commodity hardware. Because the Data Engine can access disk storage as well as RAM and cache memory, it is not limited by the amount of memory on a system. There is no requirement that an entire data set can be loaded into memory to achieve its performance goals.

4.3 Microsoft Power BI

Microsoft Power BI has long been providing BI platform technologies such as SQL Server Analysis Service (SSAS), but has long been absent from delivering client or presentation layer technologies such as Online Analytical Processing (OLAP). Excel Pivot tables have been around for a while and can facilitate simple multi-dimensional analysis, but Excel's flexibility threatens data integrity, Excel's memory limitations limit data set volumes, and this type of solution falls far short of enterprise data warehouse capabilities.

The underlying Power BI technology is an in-memory analytics engine and columnar database that supports tabular data store structures used by Power Pivot. This achieves a balance between performance and ease of use (as compared to three dimensional cubes which require more complex assembly and query languages, such as MDX (multidimensional expressions) for SSAS).

6. Tool for data visualization of QlikView, Tableau and Power BI

Comparing the tools for data visualization of Tableau and QlikView, the conclusion is that both software have many visualization tools. Standard charts (bar chart, line chart, pie chart, area chart, and scatter plot) are available in Tableau and QlikView. Additionaly, Tableau can visualize other chart such as histograms, box-and-whisker plots, filled maps, packed bubble charts and word clouds which are not available in QlikView. QlikView gives the possibility to make gauge charts, funnel charts, grid charts and Mekko chart.

Word cloud is a visualization method that displays how frequently words appear in a given body of text, by making the size of each word proportional to its frequency. In this paper as a body of text we used word from abstract and introduction of this paper, (Fig.3).

Funnel charts is often used to represent stages in a sales process and show the amount of potential revenue for each stage. This type of chart can also be useful in identifying potential problem areas in an organization's sales processes. For these reasons, this type of visualization is shown (Fig.4). The visualization shows the total revenue per country.

Power BI offers a variety of visualization options such as: bar, line, area, waterfall, treemaps, donut, pie charts, stacked charts, bubble charts, geographical charts, and gauges based on a percentage value as well as card visualization. Microsoft has made the source code for the Power BI visuals publicly available and is enabling developers to build custom visuals for Power BI.

In this paper we decided to present this type of visualization, (Fig.5) because currently is only supported in Power BI. A chart contains data for certain cities, as well as a picture of the flag of the country in which the city is located. The image is obtained through the URL that is saved in a separate column in the table.

Mapping is also integrated in the Tableau software with geospatial and thematic maps that overlay. Tableau maps do not use 3D charts and do not have movie/video recording like Power Map. Microsoft Power Map has 3D, flat thematic mapping, heatmap visualizations, data layer overlay mapping and unique movie/video recording features. Mapping is not standardly integrated in QlikView. To visualize data on a map it requires to purchase an additional extension such as GeoQlik.

Tableau and QlikView both offer the ability to do statistical analysis and forecasting, while Power BI has yet to add this capability.

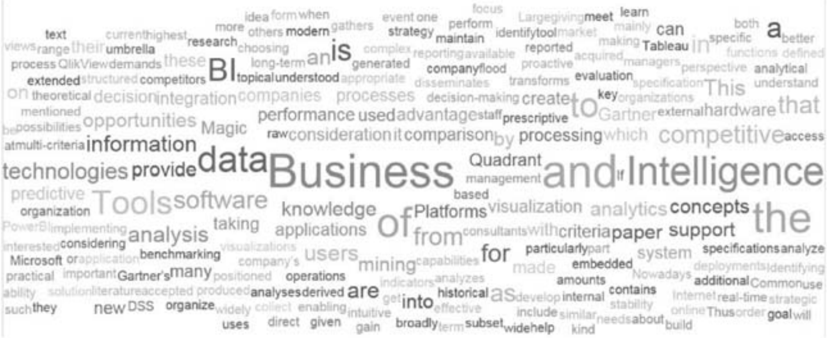
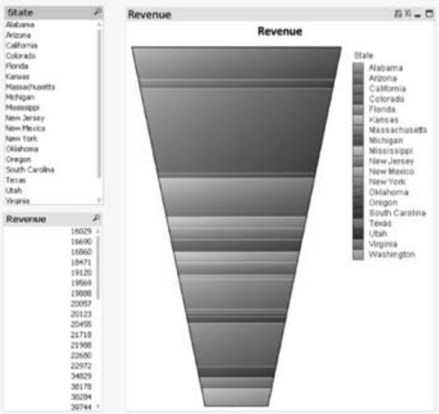


Figure 3 – Word cloud visualization created in Tableau


State
Alabama
Arizona
California
Colorado
Florida
Kansas
Massachusetts
Michigan
Mississippi
New Jersey
New Mexico
New York
Oklahoma
Oregon
South Carolina
Texas
Utah
Virginia

Revenue
16029
16690
16860
18471
19120
19569
19808
20057
20123
20495
21718
21908
22680
22972
34829
38178
38284
39744

Figure 4 – Funnel char in QlikView

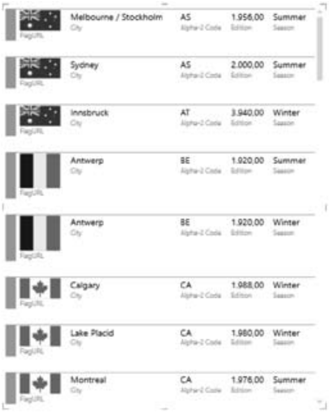

FlagURL	Melbourne / Stockholm City	AS Alpha-2 Code	1.956,00 Edition	Summer Season
FlagURL	Sydney City	AS Alpha-2 Code	2.000,00 Edition	Summer Season
FlagURL	Innsbruck City	AT Alpha-2 Code	3.940,00 Edition	Winter Season
FlagURL	Antwerp City	BE Alpha-2 Code	1.920,00 Edition	Summer Season
FlagURL	Antwerp City	BE Alpha-2 Code	1.920,00 Edition	Winter Season
FlagURL	Calgary City	CA Alpha-2 Code	1.988,00 Edition	Winter Season
FlagURL	Lake Placid City	CA Alpha-2 Code	1.980,00 Edition	Winter Season
FlagURL	Montreal City	CA Alpha-2 Code	1.976,00 Edition	Summer Season

Figure 5 – Card visualization in Power BI

Conclusion

The paper presented analysis considering business criteria as: data integration possibilities, hardware specification, comparison of used technologies and tools for data visualization taking into consideration Tableau, QlikView and Microsoft's Power BI. These software are competitive market leaders in BI, positioned in the group of leaders according to world's leading information technology research and advisory company Gartner.

From our theoretical and practical view, QlikView enable users to gain business insights by understanding how data is associated and what data is not related by in-memory architecture which addresses the problems by the traditional disk-based and query-based BI tools. Users with sufficient processing power can analyze enormous amount of data, but is not really meant for people who are not programmers.

Tableau's has user friendly drag and drop capabilities allow non-technical users to easily create and develop their dashboards. From the other side, Tableau still has weaknesses in the area of data integration across data sources. Tableau supports a diverse range of data connectivity options but offers a low level support when it comes to integrating combinations of these sources in preparation for analysis.

Microsoft Power BI offers a competitive advantage with its insightful reporting and sharing capabilities, simple visualizations, and integration into the Microsoft packages. Microsoft was ranked in the top quartile of Magic Quadrant vendors for user enablement (only Tableau ranked slightly higher), with high scores for online tutorials, community support, conferences and documentation. Microsoft capabilities for advanced analytics within Power BI are limited. Even simple forecasting must be done externally within Excel.

From our practical application for the purposes of this paper we separated three visualization: Word cloud in Tableau, Funnel cloud in QlikView and Card Visualization in Power BI.

References

[1] Sallam R.L. & all, Magic Quadrant for Business Intelligence and Analytics Platforms, Gartnet Inc., (February 2015)

[2] Evelson B. & all, The Forrester Wave™: Agile Business Intelligence Platforms, Q3 2015, for application and delivery professionals, (March 2015)

[3] Mukherjee S., Ovum Decision Matrix: Selecting a Business Intelligence Solution, 2014–15, Ovum, (July 2014), Product code: IT0014-002923

[4] Ranjan J., Business Intelligence: Concepts, Components, Technologies and Benefits, Journal of Theoretical and Applied Information Technology, Vol.9 No.1, pp.060-070, (2009)

[5] Schaefer D. & all, Delivering Self-Service BI, Data Visualization, and Big Data Analytics, IT@Intel White Paper, Intel, (June 2013)

[6] Keahey T.A., Using visualization to understand big data, IBM Corporation, (Sep 2013)

[7] [Pubdat..] Aigner W., Current Work Practice and Users' Perspectives on Visualization and Interactivity in Business Intelligence, 2013 17th International Conference on Information Visualisation, IEEE, DOI: 10.1109/IV.2013.38

[8] Chaudhuri, S., Dayal, U., & Narasayya, V. (2011). An overview of business intelligence technology. Communications of the ACM, 54(8), 88-98.

[9]https://www.google.com/search?q=Gartner+and+reported+as+the+Magic+Quadrant+2016&source=lnms&tbm=isch&sa=X&ved=0ahUKEwiXqfqTlofYAhVKDJoKHdtuBcUQ_AUICygC&biw=1600&bih=773#imgrc=tnijTFadLidsjM, Accessed 17.10.2017

[10] Dupin-Bryant P.A. and all, Business Intelligence, Analytics And Data Visualization: A Heat Map Project Tutorial, International Journal of Management & Information Systems – Third Quarter 2014 Volume 18, Number 3

[11] Thomas H. Davenport Enterprise Analytics: Optimize Performance, Process, and Decisions Through Big Data, FT Press, 13.9.2012

[12] Kandel S. et al., "Enterprise Data Analysis and Visualization: An Interview Study," IEEE Trans. Visualization and Computer Graphics, vol. 18, no. 12, 2012, pp. 2917–2926.

[13] Baltzan, P. (2014). Business driven information systems (4th Ed.). New York, NY: McGraw-Hill.

[14] Chen, H., Chiang, R. H. L., & Storey, V. C. (2012). Business intelligence and analytics: From big data to big impact. MIS Quarterly, 36(4), 1165-1188.

[15] Manyika, J. & all, Big data: The next frontier for innovation, competition, and productivity. McKinsey Global Institute: McKinsey and Company. Retrieved from http://www.mckinsey.com/Insights/MGI/Research/Technology_and_ Innovation/Big_data_The_next_frontier_for_innovation (2011), Accesed 17.3.2016

[16] Alazmi A. & Alazmi, R., Data mining and visualization of large databases. International Journal of Computer Science and Security, (2012), 6(5), 295-314.

[17] www.qlikisrael-upport.com/Knowledgebase/Article/GetAttachment/33/2091747, Accessed 15.11.2017

[18] Fisher D. et al., "Interactions with Big Data Analytics," Interactions, vol. 19, no. 3, 2012, pp. 50–59.12, pp. 45–5

[19] http://www.openskydata.com/assets/media/downloads/Power-BI-Overview-Whitepaper.pdf, Accessed 11.10.2017

[20] Savoska S., Bocevska A., Data Visualization in Business Intelligent &Analysis – Analysis of First Positioned Tools According to Gartner's Magic Quadrant in Ability to Execute, AIIT 2016, Bitola, DOI:10.20544/AIIT2016.29

Information Anonymization – Procedures, Politics and Techniques

Vladimir Dimitrov

Faculty of Mathematics and Informatics, University of Sofia,
5 James Bourchier Blvd., 1164 Sofia, Bulgaria
cht@fmi.uni-sofia.bg

Abstract. Research and investigations on computer security problems show that the most malicious problem is the information disclosure. Today this problem is enormous in the context of the new cloud services. The paper is an overview of the main computer security components: attacks, vulnerabilities and weaknesses with a focus on the last ones. An approach to information disclosure weaknesses formalization and its usage for automated weakness' discovery are discussed.

Keywords: data anonymity, information disclosure, computer security.

1 Introduction

The main terms used in computer security are attack, vulnerability and weakness.

Attack is a sequence of actions (manual, automatic or both) initiated by an attacker against a software intensive system to gain an unauthorized access to or some benefits from the system. There is always motivation for the attack. Attack targets are software weaknesses.

Software **weakness** is an intentional or unintentional bug in the system architecture, design, implementation or configuration.

Vulnerability is a weakness that can be successfully exploited by an attack. Not every weakness available in the software is a vulnerability - it can be protected by some security mechanisms.

The weaknesses represent the computer security statics of the system. The attacks represent computer security dynamic of the system. The vulnerabilities are bridge between the static and dynamic.

This conceptual system (attack, vulnerability and weakness) is supported in many computer security repositories.

"An **information exposure** is the intentional or unintentional disclosure of information to an actor that is not explicitly authorized to have access to that information." [1]

Alternative terms are **information leak** and **information disclosure**. The term "information leak" in computer security is used additionally in the sense of resource leak, i.e. improper tracking of resources which can lead to exhaustion.

The term “information disclosure” does not refer to disclosure of security-relevant information. It used mainly in vulnerability databases and policies and legal documents.

Unauthorized information disclosure implies that the information is **sensitive** in some way, i.e. it is **classified** as such one by the corresponding **authorities**. For the purpose of this paper it is not important how the information is classified as sensitive.

Data masking is a fundamental component of data security used for data **anonymization** (de-identification). It enables organizations to de-identify, mask and transform sensitive data. Anonymized data can be used for research purpose. A range of transformation techniques can be applied to substitute sensitive data with contextually-accurate but fictionalized data to produce accurate research results. By masking personally-identifying information, organizations can protect the privacy and security of confidential data, and support compliance with local and international privacy regulations.

There are 3 types of data anonymization: **masking identifiers in unstructured data**, **privacy preserving data analyses** (interactive scenario) and **transforming structured data** (non-interactive scenario). [2]

The next aspects in anonymization algorithms must be balanced: use cases, types of data, risk and thread models, privacy models (syntactic or semantic), transformation methods and utility measures (loss of information).

Important aspects of the use cases are:

- Who or what processes the data and in which way? Humans (types of analyses, interactive or non-interactive) or machines (classification, clustering).
- How will the data be released? Access control (open or restricted) or continuous data publishing (multiple views, re-release – incremental or new attributes).
- Is the data distributed? Collaborative environments (vertical or horizontal).

Important properties of the data are data type (relational or transactional), data dimensionality, and data clusterization. Transactional data consists of set-valued attributes.

Among the transformation techniques are string literal values, character substrings, random or sequential numbers, arithmetic expressions, concatenated expressions, date aging, lookup values and intelligence.

2 Motivations

Anonymization must be treated in the context of the concrete use cases. Data anonymization is data protection from unauthorized access, i.e. there is an intended or unintended attack for data disclosure. From that point of view, data disclosure is a vulnerability that occurs as result of an attack on some weakness. Therefore data anonymization is a mitigation or prevention of data disclosure weaknesses in software intensive systems.

The main subject for data anonymization are so called Personally Identifiable Information (PII) and Protected Health Information (PHI).

PII is information that alone or in combination with other information allows a person to be identified. This is information that permits individual persons to be revealed from the mass of information. Part of the PII is information associated with the person.

PII can be financial (card number, bank account number and balance on it), employment details (salary, position occupation), personal (photo, biometric data, birth date, sex, marital status), education (college, university, qualifications), contacts (e-mail, phone number), medical data (past and current diseases).

PHI is collected, generated, stored and transmitted by the health care vendors. This information directly or indirectly identifies the individuals.

Some of the reasons for data anonymizations:

1. The business generates sensitive data and these data must be protected.
2. The malicious use of personal data is subject of regulatory and legal prosecutions.
3. The misuse of sensitive data generate enormous loses for the business.
4. There are operational risks in the context of outsourcing and cooperation.
5. There are legal and compliance requirements.

The main goal of the current research is to identify the weaknesses that can gain to data disclosure. For that purpose Common Weaknesses Enumeration (CWE) [3] is used.

3 Weaknesses

“CWE-200: Information Exposure” is a class weakness. It status is incomplete. It participates in many views and is starting point for this research. CWE-200 is the ancestor of the next weaknesses:

- CWE-201: Information Exposure Through Sent Data (variant) (draft)
- CWE-202: Exposure of Sensitive Data Through Data Queries (variant) (draft)
- CWE-203: Information Exposure Through Discrepancy (class) (incomplete)
    - CWE-204: Response Discrepancy Information Exposure (base) (incomplete)
    - CWE-205: Information Exposure Through Behavioral Discrepancy (base) (incomplete)
        - CWE-206: Information Exposure of Internal State Through Behavioral Inconsistency (variant) (incomplete)

        - CWE-207: Information Exposure Through an External Behavioral Inconsistency (variant) (draft)
    - CWE-208: Information Exposure Through Timing Discrepancy (base) (incomplete)
- CWE-209: Information Exposure Through an Error Message (base) (draft)
    - CWE-210: Information Exposure Through Self-generated Error Message (base) (draft)
        - CWE-535: Information Exposure Through Shell Error Message (variant) (incomplete)
        - CWE-536: Information Exposure Through Servlet Runtime Error Message (variant) (incomplete)
        - CWE-537: Information Exposure Through Java Runtime Error Message (variant) (incomplete)
    - CWE-211: Information Exposure Through Externally-generated Error Message (base) (incomplete)
    - CWE-550: Information Exposure Through Server Error Message (variant) (incomplete)
- CWE-212: Improper Cross-boundary Removal of Sensitive Data (base) (incomplete)
- CWE-213: Intentional Information Exposure (base) (draft)
- CWE-214: Information Exposure Through Process Environment (variant) (incomplete)
- CWE-215: Information Exposure Through Debug Information (variant) (draft)
    - CWE-11: ASP.NET Misconfiguration: Creating Debug Binary (variant) (draft)
- CWE-226: Sensitive Information Uncleared Before Release (base) (draft)
    - CWE-244: Improper Clearing of Heap Memory Before Release (‘Heap Inspection’) (variant) (draft)
- CWE-359: Exposure of Private Information (‘Privacy Violation’) (class) (incomplete)
    - CWE-202: Exposure of Sensitive Data Through Data Queries (variant) (draft)
- CWE-497: Exposure of System Data to an Unauthorized Control Sphere (variant) (incomplete)
- CWE-524: Information Exposure Through Caching (variant) (incomplete)
    - CWE-525: Information Exposure Through Browser Caching (variant) (incomplete)
- CWE-526: Information Exposure Through Environmental Variables (variant) (incomplete)

- CWE-538: File and Directory Information Exposure (base) (draft)
    - CWE-527: Exposure of CVS Repository to an Unauthorized Control Sphere (variant) (draft)
    - CWE-528: Exposure of Core Dump File to an Unauthorized Control Sphere (variant) (draft)
    - CWE-529: Exposure of Access Control List Files to an Unauthorized Control Sphere (variant) (incomplete)
    - CWE-530: Exposure of Backup File to an Unauthorized Control Sphere (variant) (incomplete)
    - CWE-532: Information Exposure Through Log Files (variant) (incomplete)
        - CWE-533: Information Exposure Through Server Log Files (variant) (incomplete)
        - CWE-534: Information Exposure Through Debug Log Files (variant) (draft)
        - CWE-542: Information Exposure Through Cleanup Log Files (variant) (incomplete)
    - CWE-539: Information Exposure Through Persistent Cookies (variant) (incomplete)
    - CWE-540: Information Exposure Through Source Code (variant) (incomplete)
        - CWE-531: Information Exposure Through Test Code (variant) (incomplete)
        - CWE-541: Information Exposure Through Include Source Code (variant) (incomplete)
        - CWE-615: Information Exposure Through Comments (variant) (incomplete)
    - CWE-548: Information Exposure Through Directory Listing (variant) (draft)
    - CWE-651: Information Exposure Through WSDL File (variant) (incomplete)
- CWE-598: Information Exposure Through Query Strings in GET Request (variant) (draft)
- CWE-612: Information Exposure Through Indexing of Private Data (variant) (draft)

Now, let's analyze this hierarchy of weaknesses.

Every CWE node can be in one of the next statuses:

- "stable" means that the node is ready for public discussion.
- "draft" means that only an initial description is available and further clarification must be done.
- "incomplete" means that the node is not carefully investigated in details.

The life cycle of the node is draft-incomplete-stable. In the last state, the node can be fully utilized.

First, all these weaknesses are in status "incomplete" or "draft". In that case, how useful is the presented information for further research? It is clear that with information in such a state is impossible to be done more formal research. The hierarchy is under development and only after one or two new versions would be in status dominated by "stable" states.

Second, the hierarchy is organized in very strange way. Today, object-oriented approach dominates in the programming and computer professionals can expect that hierarchy is organized following the logic class-base-variant, i.e. the most specific are variants that can be generalized to bases and the last one can be generalized to classes. On the top are views - in that case the Research View.

Several words about views, classes, bases and variants:

- The view organizes weaknesses in a hierarchy from specific point of view.
- The class describes a set of weaknesses with common characteristics. It description usually is independent of specific programming language or technology.
- The base is an abstract description more specific than the class, but it can be used for weakness detection and prevention.
- The variant has the most detailed description linked with some programming language or technology.

CWE-200 hierarchy contains nodes of the same type in relationship parent-child. For example CWE-532 (variant) has as children the variants CWE-533, CWE-534 and CWE-542.

In that hierarchy nodes of intermediate types are not obligatory. For Example, CWE-201 (variant) directly is a child of CWE-200 (class).

The generalization, in object-oriented programming and design, uses a single concept to generalize more specific elements in one more abstract one. In the CWE case, this means that several variants have to be generalized in one base, and that several bases have to be generalized in one class because they are at different abstraction levels. But from the CWE-200 hierarchy it is clear that the concepts of class, base and variant are not used for that purpose. For example CWE-200 have as children several variants (CWE-201 etc.), a class (CWE-203), and several bases (CWE-209 etc.).

The concepts of class, base and variant are used as node attribute describing its abstraction level and they have no relation with hierarchy organization.

Now, the question is "How CWE-200 hierarchy is organized?" After careful investigation of the contents it becomes clear that the hierarchy is organized following attack vectors hierarchy.

An attack vector is a path or means by which an attacker can gain access to the system. Attack vectors enable attackers to exploit system vulnerabilities.

Attack vectors at the top of CWE-200 hierarchy are:

- data transmission (CWE-201);
- query execution (CWE-202);
- discrepancies (CWE-203):
    - response discrepancies (CWE-204);
    - behavior discrepancies (CWE-205):
        - internal state behavioral inconsistency (CWE-206);
        - external state behavioral inconsistency (CWE-207).
    - timing discrepancies (CWE-208).
- error messages (CWE-209);
- etc.

Third, CWE-202 participates two times in the hierarchy: one time as a child of CWE-200 and the second time as a child of CWE-359. This hierarchy is at least strange.

A hierarchy of weaknesses must help to detect them in the code. It must be a base for further research and investigation on these weaknesses. This means that the hierarchy must be organized in object-oriented style - from more abstract to more specific elements. Every node in this hierarchy must be self-contained and useful, i.e. its formal specification must not be trivial one. If this does not happened - the node contains pointless description. The node specification must be the base for automatic detection of the corresponding weakness. The description must not be a common parlance on the topic.

In the Appendix is given a summary of weaknesses that can gain to information disclosure.

4 Mitigations

Information disclosure can be established during the system architecture, design or implementation phases. It is not related with specific programming languages or technologies. This class of weaknesses more frequently can be found for the architecture paradigm of mobile applications.

Consequences from the successful exploitation of information disclosure weaknesses are on confidentiality because the attackers can read application data.

This class of weaknesses can be detected using the following methods:

- automatic static analyses of binary code or bytecode with partial efficiency;
- dynamic analyses with automatic interpretation of the results with high efficiency;
- dynamic analyses with manual interpretation of the results with partial efficiency;
- manual static analyses of source code with high efficiency;

- automatic static analyses of source code with high efficiency;
- architecture or design review with high efficiency.

This class of weaknesses can be mitigated during architecture / design phase using the principle of privilege separation. The system must be partitioned in safe areas of trust. Sensitive data must not cross safe area boundaries. The system must be built on these safe areas. The privileges must be managed by the principle of least privilege, i.e. the trusted entity should receive only the privilege needed to perform its operations and when the need of such a privilege is way it must be dropped.

5 Conclusion

Data anonymization is the perfect solution for data protection, because even in the case of successful attack, the attacker access useless data, but there are two problems with data anonymization:

1. How useful for data analysis are anonymized data?
2. The process for data anonymization still remains a subject for data disclosure attacks. Therefore, it must be protected enough for that kind of attacks using above mentioned weaknesses.

The idea to use software weaknesses for data security improvement is a good idea - a preventive action. It is possible to prevent weaknesses as early as possible in the software life cycle. Later on, to remove discovered vulnerabilities is expensive or even impossible.

On the other hand, hierarchy of information disclosure weaknesses is currently under development - there are no one weakness in stable state. This hierarchy is only informative one and as a whole not usable. Will the hierarchy be usable when it reaches the stable state is under question that would be discussed in another place.

It is clear that to protect data from disclosure using CWE weaknesses would not happened for now. Then what to do? An answer is to use available repositories for attacks and vulnerabilities. Every important technology or product has such repositories.

There are no links among these specific repositories. It is possible a vulnerability reported in one repository (and even currently removed) to be available for the same kind of technology or product (and even not reported). Typical example is CVE-2002-2031 that has been reported for Internet Explorer and is not available for the newest versions, but is in full power and even not reported for all other widespread browsers.

Acknowledgments. Presented in this paper results are part of the NSF project “GloBIG: A Model of Integration of Cloud Framework for Hybrid Massive Parallelism and its Application for Analysis and Automated Semantic

Enhancement of Big Heterogeneous Data Collections”, Contract ДН 02/9 of 17.12.2016.

References

1. MITRE Corp., CWE-200: Information Exposure, http://cwe.mitre.org/data/definitions/200.html (visited on 07/30/2017)
2. ARX, ARX – Powerful Data Anonymization, http://arx.deidentifier.org (visited on 07/30/2017)
3. MITRE Corp., Common Weaknesses Enumeration, http://cwe.mitre.org (visited on 07/30/2017)

Appendix: CWEs

A.1 CWE-201: Information Exposure Through Sent Data

The accidental exposure of sensitive information through sent data refers to the transmission of data which are either sensitive in and of itself or useful in the further exploitation of the system through standard data channels.

A.2 CWE-202: Exposure of Sensitive Data Through Data Queries

When trying to keep information confidential, an attacker can often infer some of the information by using statistics.

In situations where data should not be tied to individual users, but a large number of users should be able to make queries that “scrub” the identity of users, it may be possible to get information about a user -- e.g., by specifying search terms that are known to be unique to that user.

A.3 CWE-203 Information Exposure Through Discrepancy

The product behaves differently or sends different responses in a way that exposes security-relevant information about the state of the product, such as whether a particular operation was successful or not.

A.3.1 CWE-204: Response Discrepancy Information Exposure

The software provides different responses to incoming requests in a way that allows an actor to determine system state information that is outside of that actor’s control sphere.

This issue frequently occurs during authentication, where a difference in failed-login messages could allow an attacker to determine if the username is valid or not. These exposures can be inadvertent (bug) or intentional (design).

A.3.2 CWE-205: Information Exposure Through Behavioral Discrepancy

The product’s actions indicate important differences based on (1) the internal state of the product or (2) differences from other products in the same class.

For example, attacks such as OS fingerprinting rely heavily on both behavioral and response discrepancies.

A.3.2.1 CWE-206: Information Exposure of Internal State Through Behavioral Inconsistency

Two separate operations in a product cause the product to behave differently in a way that is observable to an attacker and reveals security-relevant information about the internal state of the product, such as whether a particular operation was successful or not.

A.3.2.2 CWE-207: Information Exposure Through an External Behavioral Inconsistency

“The product behaves differently than other products like it, in a way that is observable to an attacker and exposes security-relevant information about which product is being used.”

A.3.3 CWE-208: Information Exposure Through Timing Discrepancy

Two separate operations in a product require different amounts of time to complete, in a way that is observable to an actor and reveals security-relevant information about the state of the product, such as whether a particular operation was successful or not.

A.4 CWE-209: Information Exposure Through an Error Message

The software generates an error message that includes sensitive information about its environment, users, or associated data.

The sensitive information may be valuable information on its own (such as a password), or it may be useful for launching other, more deadly attacks. If an attack fails, an attacker may use error information provided by the server to launch another more focused attack. For example, an attempt to exploit a path traversal weakness (CWE-22) might yield the full pathname of the installed application. In turn, this could be used to select the proper number of “..” sequences to navigate to the targeted file. An attack using SQL injection (CWE-89) might not initially succeed, but an error message could reveal the malformed query, which would expose query logic and possibly even passwords or other sensitive information used within the query.

A.4.1 CWE-210: Information Exposure Through Self-generated Error Message

The software identifies an error condition and creates its own diagnostic or error messages that contain sensitive information.

A.4.1.1 CWE-535: Information Exposure Through Shell Error Message

A command shell error message indicates that there exists an unhandled exception in the web application code. In many cases, an attacker can leverage the conditions that cause these errors in order to gain unauthorized access to the system.

A.4.1.2 CWE-536: Information Exposure Through Servlet Runtime Error Message

A servlet error message indicates that there exists an unhandled exception in your web application code and may provide useful information to an attacker.

A.4.1.3 CWE-537: Information Exposure Through Java Runtime Error Message

In many cases, an attacker can leverage the conditions that cause unhandled exception errors in order to gain unauthorized access to the system.

A.4.2 CWE-211: Information Exposure Through Externally-generated Error Message

The software performs an operation that triggers an external diagnostic or error message that is not directly generated by the software, such as an error generated by the programming language interpreter that the software uses. The error can contain sensitive system information.

A.4.3 CWE-550: Information Exposure Through Server Error Message

Certain conditions, such as network failure, will cause a server error message to be displayed.

While error messages in and of themselves are not dangerous, per se, it is what an attacker can glean from them that might cause eventual problems.

A.5 CWE-212: Improper Cross-boundary Removal of Sensitive Data

The software uses a resource that contains sensitive data, but it does not properly remove that data before it stores, transfers, or shares the resource with actors in another control sphere.

Resources that may contain sensitive data include documents, packets, messages, databases, etc. While this data may be useful to an individual user or small set of users who share the resource, it may need to be removed before the resource can be shared outside of the trusted group. The process of removal is sometimes called cleansing or scrubbing.

For example, software that is used for editing documents might not remove sensitive data such as reviewer comments or the local pathname where the document is stored. Or, a proxy might not remove an internal IP address from headers before making an outgoing request to an Internet site.

A.6 CWE-213: Intentional Information Exposure

A product's design or configuration explicitly requires the publication of information that could be regarded as sensitive by an administrator.

A.7 CWE-214: Information Exposure Through Process Environment

A process is invoked with sensitive arguments, environment variables, or other elements that can be seen by other processes on the operating system.

Many operating systems allow a user to list information about processes that are owned by other users. This information could include command line arguments or environment variable settings. When this data contains sensitive information such as credentials, it might allow other users to launch an attack against the software or related resources.

A.8 CWE-215: Information Exposure Through Debug Information

The application contains debugging code that can expose sensitive information to untrusted parties.

A.8.1 CWE-11: ASP.NET Misconfiguration: Creating Debug Binary

Debugging messages help attackers learn about the system and plan a form of attack.
ASP .NET applications can be configured to produce debug binaries. These binaries give detailed debugging messages and should not be used in production environments. Debug binaries are meant to be used in a development or testing environment and can pose a security risk if they are deployed to production.

A.9 CWE-226: Sensitive Information Uncleared Before Release

The software does not fully clear previously used information in a data structure, file, or other resource, before making that resource available to a party in another control sphere.

This typically results from new data that is not as long as the old data, which leaves portions of the old data still available. Equivalent errors can occur in other situations where the length of data is variable but the associated data structure is not. If memory is not cleared after use, it may allow unintended actors to read the data when the memory is reallocated.

A.9.1 CWE-244: Improper Clearing of Heap Memory Before Release ('Heap Inspection')

Using realloc() to resize buffers that store sensitive information can leave the sensitive information exposed to attack, because it is not removed from memory.

When sensitive data such as a password or an encryption key is not removed from memory, it could be exposed to an attacker using a “heap inspection” attack that reads the sensitive data using memory dumps or other methods. The realloc() function is commonly used to increase the size of a block of allocated memory. This operation often requires copying the contents of the old memory block into a new and larger block. This operation leaves the contents of the original block intact but inaccessible to the program, preventing the program from being able to scrub sensitive data from memory. If an attacker can later examine the contents of a memory dump, the sensitive data could be exposed.

A.10 CWE-359: Exposure of Private Information (‘Privacy Violation’)

The software does not properly prevent private data (such as credit card numbers) from being accessed by actors who either (1) are not explicitly authorized to access the data or (2) do not have the implicit consent of the people to which the data is related.

Mishandling private information, such as customer passwords or Social Security numbers, can compromise user privacy and is often illegal. An exposure of private information does not necessarily prevent the software from working properly, and in fact it might be intended by the developer, but it can still be undesirable (or explicitly prohibited by law) for the people who are associated with this private information.

Privacy violations may occur when:

1. Private user information enters the program.
2. The data is written t an external location, such as the console, file system, or network.

Private data can enter a program in a variety of ways:

1. Directly from the user in the form of a password or personal information
2. Accessed from a database or other data store by the application
3. Indirectly from a partner or other third party

Some types of private information include:

- Government identifiers, such as Social Security Numbers
- Contact information, such as home addresses and telephone numbers
- Geographic location - where the user is (or was)
- Employment history
- Financial data - such as credit card numbers, salary, bank accounts, and debts
- Pictures, video, or audio
- Behavioral patterns - such as web surfing history, when certain activities are performed, etc.
- Relationships (and types of relationships) with others - family, friends, contacts, etc.

- Communications - e-mail addresses, private e-mail messages, SMS text messages, chat logs, etc.
- Health - medical conditions, insurance status, prescription records
- Credentials, such as passwords, which can be used to access other information.

Some of this information may be characterized as PII (Personally Identifiable Information), Protected Health Information (PHI), etc. Categories of private information may overlap or vary based on the intended usage or the policies and practices of a particular industry.

Depending on its location, the type of business it conducts, and the nature of any private data it handles, an organization may be required to comply with one or more of the following federal and state regulations: - Safe Harbor Privacy Framework [R.359.2] - Gramm-Leach Bliley Act (GLBA) [R.359.3] - Health Insurance Portability and Accountability Act (HIPAA) [R.359.4] - California SB-1386 [R.359.5].

Sometimes data that is not labeled as private can have a privacy implication in a different context. For example, student identification numbers are usually not considered private because there is no explicit and publicly-available mapping to an individual student's personal information. However, if a school generates identification numbers based on student social security numbers, then the identification numbers should be considered private.

Security and privacy concerns often seem to compete with each other. From a security perspective, all important operations should be recorded so that any anomalous activity can later be identified. However, when private data is involved, this practice can in fact create risk. Although there are many ways in which private data can be handled unsafely, a common risk stems from misplaced trust. Programmers often trust the operating environment in which a program runs, and therefore believe that it is acceptable store private information on the file system, in the registry, or in other locally-controlled resources. However, even if access to certain resources is restricted, this does not guarantee that the individuals who do have access can be trusted.

A.11 CWE-497: Exposure of System Data to an Unauthorized Control Sphere

Exposing system data or debugging information helps an adversary learn about the system and form an attack plan.

An information exposure occurs when system data or debugging information leaves the program through an output stream or logging function that makes it accessible to unauthorized parties. An attacker can also cause errors to occur by submitting unusual requests to the web application. The response to these errors can reveal detailed system information, deny service, cause security mechanisms to fail, and crash the server. An attacker can use error messages that reveal technologies, operating systems, and product versions to tune the attack against known vulnerabilities in these technologies. An application may use diagnostic methods that provide significant implementation details such as stack traces as part of its error handling mechanism.

A.12 CWE-524: Information Exposure Through Caching

The application uses a cache to maintain a pool of objects, threads, connections, pages, or passwords to minimize the time it takes to access them or the resources to which they connect. If implemented improperly, these caches can allow access to unauthorized information or cause a denial of service vulnerability.

A.12.1 CWE-525: Information Exposure Through Browser Caching

For each web page, the application should have an appropriate caching policy specifying the extent to which the page and its form fields should be cached.

A.13 CWE-526: Information Exposure Through Environmental Variables

Environmental variables may contain sensitive information about a remote server.

A.14 CWE-538: File and Directory Information Exposure

The product stores sensitive information in files or directories that are accessible to actors outside of the intended control sphere.

A.14.1 CWE-527: Exposure of CVS Repository to an Unauthorized Control Sphere

The product stores a CVS repository in a directory or other container that is accessible to actors outside of the intended control sphere.

Information contained within a CVS subdirectory on a web server or other server could be recovered by an attacker and used for malicious purposes. This information may include usernames, filenames, path root, and IP addresses.

A.14.2 CWE-528: Exposure of Core Dump File to an Unauthorized Control Sphere

The product generates a core dump file in a directory that is accessible to actors outside of the intended control sphere.

A.14.3 CWE-529: Exposure of Access Control List Files to an Unauthorized Control Sphere

The product stores access control list files in a directory or other container that is accessible to actors outside of the intended control sphere.

Exposure of these access control list files may give the attacker information about the configuration of the site or system. This information may then be used to bypass the intended security policy or identify trusted systems from which an attack can be launched.

A.14.4 CWE-530: Exposure of Backup File to an Unauthorized Control Sphere

A backup file is stored in a directory that is accessible to actors outside of the intended control sphere.

Often, old files are renamed with an extension such as .~bk to distinguish them from production files. The source code for old files that have been renamed in this manner and left in the webroot can often be retrieved. This renaming may have been performed automatically by the web server, or manually by the administrator.

A.14.5 CWE-532: Information Exposure Through Log Files

Information written to log files can be of a sensitive nature and give valuable guidance to an attacker or expose sensitive user information.

While logging all information may be helpful during development stages, it is important that logging levels be set appropriately before a product ships so that sensitive user data and system information are not accidentally exposed to potential attackers.

A.14.5.1 CWE-533: Information Exposure Through Server Log Files

A server.log file was found. This can give information on whatever application left the file. Usually this can give full path names and system information, and sometimes usernames and passwords.

A.14.5.2 CWE-534: Information Exposure Through Debug Log Files

The application does not sufficiently restrict access to a log file that is used for debugging.

A.14.5.3 CWE-542: Information Exposure Through Cleanup Log Files

The application does not properly protect or delete a log file related to cleanup.

A.14.6 CWE-539: Information Exposure Through Persistent Cookies

Persistent cookies are cookies that are stored on the browser’s hard drive. This can cause security and privacy issues depending on the information stored in the cookie and how it is accessed.

Cookies are small bits of data that are sent by the web application but stored locally in the browser. This lets the application use the cookie to pass information between pages and store variable information. The web application controls what information is stored in a cookie and how it is used. Typical types of information stored in cookies are session Identifiers, personalization and customization information, and in rare cases even usernames to enable automated logins. There are two different types of cookies: session cookies and persistent cookies. Session cookies just live in the browser's memory, and are not stored anywhere, but persistent cookies are stored on the browser's hard drive.

A.14.7 CWE-540: Information Exposure Through Source Code

Source code on a web server often contains sensitive information and should generally not be accessible to users.

There are situations where it is critical to remove source code from an area or server. For example, obtaining Perl source code on a system allows an attacker to understand the logic of the script and extract extremely useful information such as code bugs or logins and passwords.

A.14.7.1 CWE-531: Information Exposure Through Test Code

Accessible test applications can pose a variety of security risks. Since developers or administrators rarely consider that someone besides themselves would even know about the existence of these applications, it is common for them to contain sensitive information or functions.

A.14.7.2 CWE-541: Information Exposure Through Include Source Code

If an include file source is accessible, the file can contain usernames and passwords, as well as sensitive information pertaining to the application and system.

A.14.7.3 CWE-615: Information Exposure Through Comments

While adding general comments is very useful, some programmers tend to leave important data, such as: filenames related to the web application, old links or links which were not meant to be browsed by users, old code fragments, etc.

An attacker who finds these comments can map the application's structure and files, expose hidden parts of the site, and study the fragments of code to reverse engineer the application, which may help develop further attacks against the site.

A.14.8 CWE-548: Information Exposure Through Directory Listing

A directory listing is inappropriately exposed, yielding potentially sensitive information to attackers.

A directory listing provides an attacker with the complete index of all the resources located inside of the directory. The specific risks and consequences vary depending on which files are listed and accessible.

A.14.9 CWE-651: Information Exposure Through WSDL File

The Web services architecture may require exposing a WSDL file that contains information on the publicly accessible services and how callers of these services should interact with them (e.g. what parameters they expect and what types they return).

An information exposure may occur if any of the following apply:

1. The WSDL file is accessible to a wider audience than intended.
2. The WSDL file contains information on the methods/services that should not be publicly accessible or information about deprecated methods. This problem is made more likely due to the WSDL often being automatically generated from the code.
3. Information in the WSDL file helps guess names/locations of methods/ resources that should not be publicly accessible.

A.15 CWE-598: Information Exposure Through Query Strings in GET Request

The web application uses the GET method to process requests that contain sensitive information, which can expose that information through the browser's history, Referers, web logs, and other sources.

A.16 CWE-612: Information Exposure Through Indexing of Private Data

The product performs an indexing routine against private documents, but does not sufficiently verify that the actors who can access the index also have the privileges to access the private documents.

When an indexing routine is applied against a group of private documents, and that index's results are available to outsiders who do not have access to those documents, then outsiders might be able to obtain sensitive information by conducting targeted searches. The risk is especially dangerous if search results include surrounding text that was not part of the search query. This issue can appear in search engines that are not configured (or implemented) to ignore critical files that should remain hidden; even without permissions to download these files directly, the remote user could read them.

Object Oriented Analysis and Design of Protein Fingerprints

Monica Dimitrova[*], Dimitar Vassilev

Faculty of Mathematics and Informatics, Sofia University “St. Kliment Ohridski”,
5 James Bourchier blvd., Sofia 1164, Bulgaria
* Corresponding author: monicabdimitrova@gmail.com

Abstract. For better analysis of amino acids and subsequent protein structures are used certain categories as motifs and fingerprints. A motif is a short, conservative segment of the amino acid sequence, whilst the fingerprint is a set of ordered motifs. These categories describe the functionality of proteins by providing more exact classification. For development of a protein fingerprint search algorithm, an object oriented model is suggested. It is based on the concept of objects containing data in the form of fields called attributes and functionalities as methods that can access and modify the data of the objects they are associated with. The major of the work includes the following tasks: comparison of fingerprints and motifs, assessment of matches, statistical significance estimation and visualization. These tasks are implemented in a web application based on data from the PRINTS database.

Keywords: protein fingerprints, pattern recognition, object-oriented analysis, bioinformatics, databases

1 Introduction

Nowadays technology is developing at a very fast pace which affects science and in particular bioinformatics. The amount of data retrieved from various sequencing projects has exploded in the last two decades and it has become very important to structure and categorize this data in order to analyze and use it in a rational manner. The key instrument in achieving this goal is the continuous improvement of existing methods and software that implements them as well as the developing of new approaches. One of the most common problems in bioinformatics is understanding the relationship between amino acid sequences and three-dimensional structure of the proteins. This relationship is not simple but much progress has been made in categorizing proteins based on their sequences, and this knowledge is used in protein modeling.

The present methods used for in silico inference of gene function rely mostly on the identification of relationships between novel sequences and those of known function. The similarity found at the sequence level is assumed to be reflected by similarity at the levels of function and structure. The analysis of uncharacterized proteins usually consists of scanning the full sequence against one or more databases that are available publicly[7]. Databases are divided into two categories - primary data sources, e.g. SWISS-PROT (Bairoch and Apweiler) [3], OWL (Web Ontology Language - Bleasby, Akrigg, Attwood), and secondary data sources[2] that condense the information from the primary databases into more and different potent identifiers (motifs, profiles, etc.) of evolutionary relationships, such as PROSITE (Bairoch et al.), BLOCKS (Henikoff et al.) and PRINTS (Attwood et al.). Such databases store reduced descriptions of protein families and can be practically used for predicting the functions and structures of novel proteins[7].

1. 1. Aim of the study and tasks

This study aims to analyze the methodology and modernize the approach for searching of protein fingerprints in PRINTS - a widely applicable bioinformatic resource. To achieve this, the current software tool will be reworked while preserving its functionality and building a new web application that will provide a better solution and easier future development. By fulfilling these goals, PRINTS will become even more important in the field of bioinformatics and will be able to evolve alongside modern IT technologies. In order to achieve the stated goals, the following tasks should be completed:

1) Researching PRINTS and the algorithm for searching of protein fingerprints. A short introduction to the specific bioinformatics theory will be presented for this purpose.

2) Developing a console application that finds the best matches in the PRINTS database on a given amino acid sequence.

3) Integrating the application results with external bioinformatics tools

2 Protein fingerprints

The two categories of methods for identifying proteins use either profiles or motifs (short fragments of sequences). The former approach is based on the compiling of a familial discriminator that contains both conserved and non-conserved regions of a multiple sequence alignment. In comparison, the motif approaches extract only the most conserved regions and can be divided into those that use a single motif and those that utilize multiple motifs. The single-motif methods, however, do not offer a biological context since only one conserved region is not enough for a match and might miss distant relatives that contain a vague version of the pattern.

The PRINTS database uses multiple conserved motifs in order to create signatures that correspond to family memberships[8]. It is usual to find more than one motif belonging to a protein family within a multiple sequence alignment and as more motifs are used, the matching with their natural neighbours increases.

A set of such motifs is defined as a fingerprint and is highly informative for the identification of distant relatives in a database search - mismatches are tolerated both at individual residues level and at motif level. Usually, the motifs are separated along the sequence and do not overlap, however they may be contiguous in 3D-space[2].

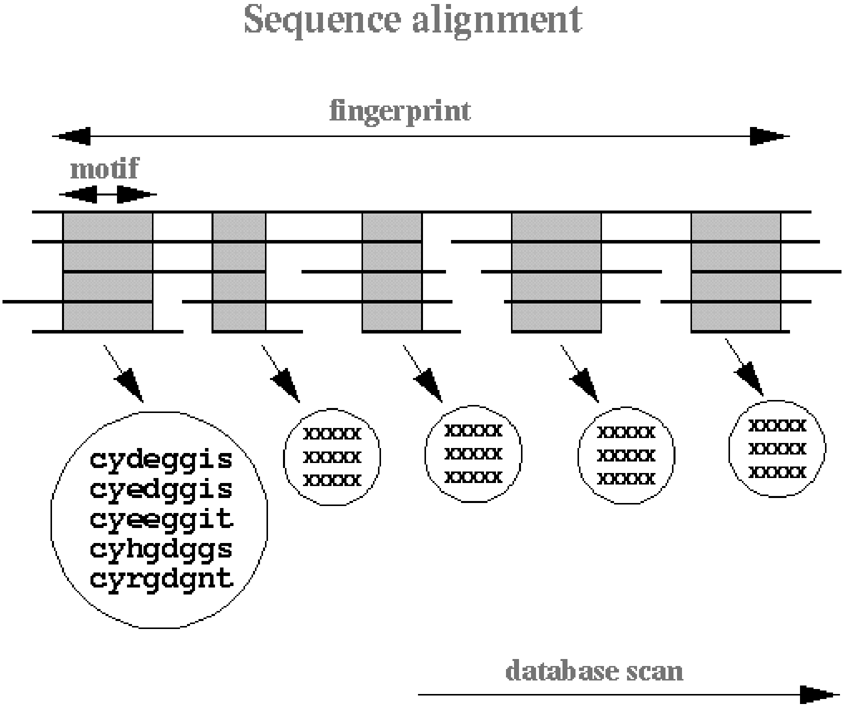


Figure 1 - generation of protein fingerprints with their corresponding motifs[10]

2.1. PRINTS database structure and functionaity

PRINTS is an important information resource for bioinformatics that was created at the end of the last century and is constantly being developed. It contributes unique functionality to InterPro and offers a range of new analysis and tools used in the annotation process. PRINTS improves the quality of sequencing because novel proteins can be compared with the whole database or particular sequences in order to examine their structures and functionalities. With this knowledge familial hierarchies can be made explicit and associations can be traced from subfamily, through family, to superfamily relations[1].

There are two kinds of fingerprints presented in the database depending on their complexity - simple and composite. Simple fingerprints are essentially single motives, while composite fingerprints encode multiple motives[9]. The majority of the database records are of the second type because the possibility of differentiation is greater in a search for many components and the results are easier to interpret.

The evolution and development of PRINTS makes sense and is possible thanks to the cooperation with other databases and projects. One of the most recent and notable projects is the integration of PRINTS with InterPro and the resolving of protein family memberships as effective as possible in order to help InterPro's automated sequence analysis. Another successful collaboration was the European Kidney and Urine Proteomics project (EuroKUP) in which a range of medically relevant protein families were studied to build hierarchical fingerprints for families in order to gain a better understanding of specific sequence properties that might cause chronic kidney diseases[1].

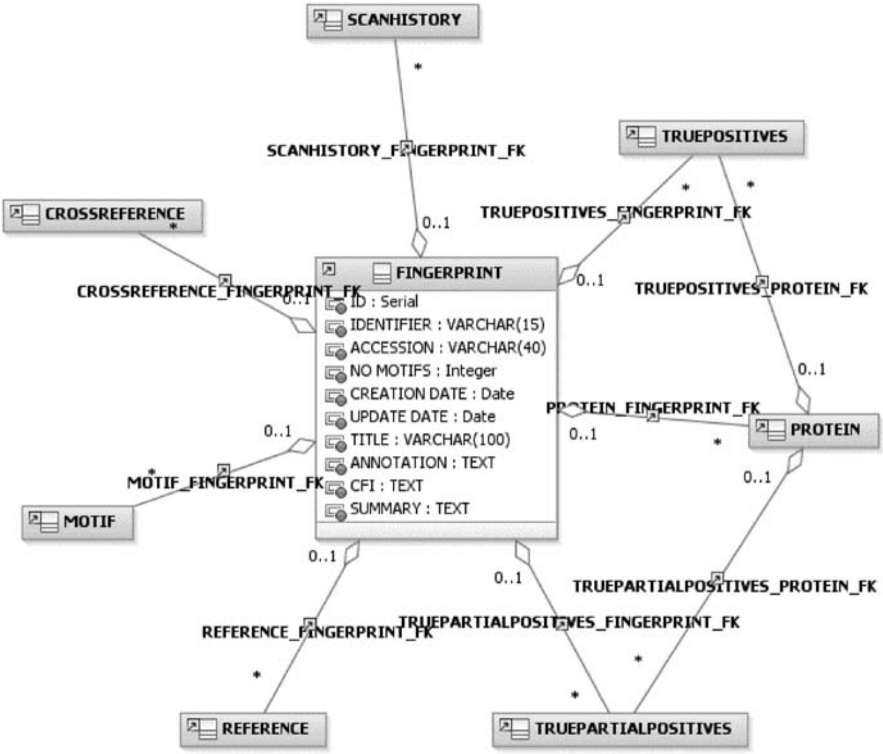


Figure 2 - UML diagram of PRINTS[9]

In order to continue developing, PRINTS and the related software should be updated regularly. Originally, PRINTS was built as a single ASCII text file[2]. This type of storing is quite common amongst molecular biology sources created in the past. However, it's not practical anymore because working with such data is unproductive especially when communicating with other databases. Relational databases have become really popular and widely used due to the speed and convenience of adding, deleting and modifying of records. Recently, a relational database has been created from the information in PRINTS in order to facilitate further development of both new tools and the database itself[9]. The information was logically separated while keeping the normalization practices and conventions.

As illustrated in Figure 2, the main table in PRINTS is **FINGERPRINT**. It is used to describe a certain fingerprint and the most valuable fields are the ID, title, annotation and set of motifs that belong to the fingerprint. Another highly used table is **MOTIF** - every record is a representation of a motif with the corresponding title, code, length and position in the fingerprint. Every motif has a set of sequences (variants), obtained in a multiple sequence alignment, described in the table **SEQ**. This table contains a certain sequence, code, start position and interval. These three tables provide an extended view of the fingerprints and contain the information needed to perform a search of an uncharacterized protein against PRINTS.

2.2. Algorithm for searching

As the database has been modernized it is desirable to update the associated tools, such as FingerPRINTScan that is used for searches against fingerprints in order to provide a diagnostic identification. This tool searches a given sequence of amino acids against PRINTS to identify the best or closest match. This information can be used for indication of the family to which the unknown protein belongs.

The algorithm for searching, illustrated in Figure 3, compares the query sequence against every fingerprint and finds the top results. In order to score a fingerprint all of its motifs should be considered - not every motif should be a match but the order must be preserved. Frequency tables and motif profiles[7] are generated for every motif based on their variants and only the motifs with scores higher than the query threshold are reported as matches. This algorithm allows identification of the best matching fingerprint to a query sequence, relying both on scores and biological information[7].

The modernization of the tool will provide a completely new code and web interface that follow the tendencies in technology while keeping the original algorithm in order to provide the expected results. The project will be open source and this way future maintenance and improvements will be easier.

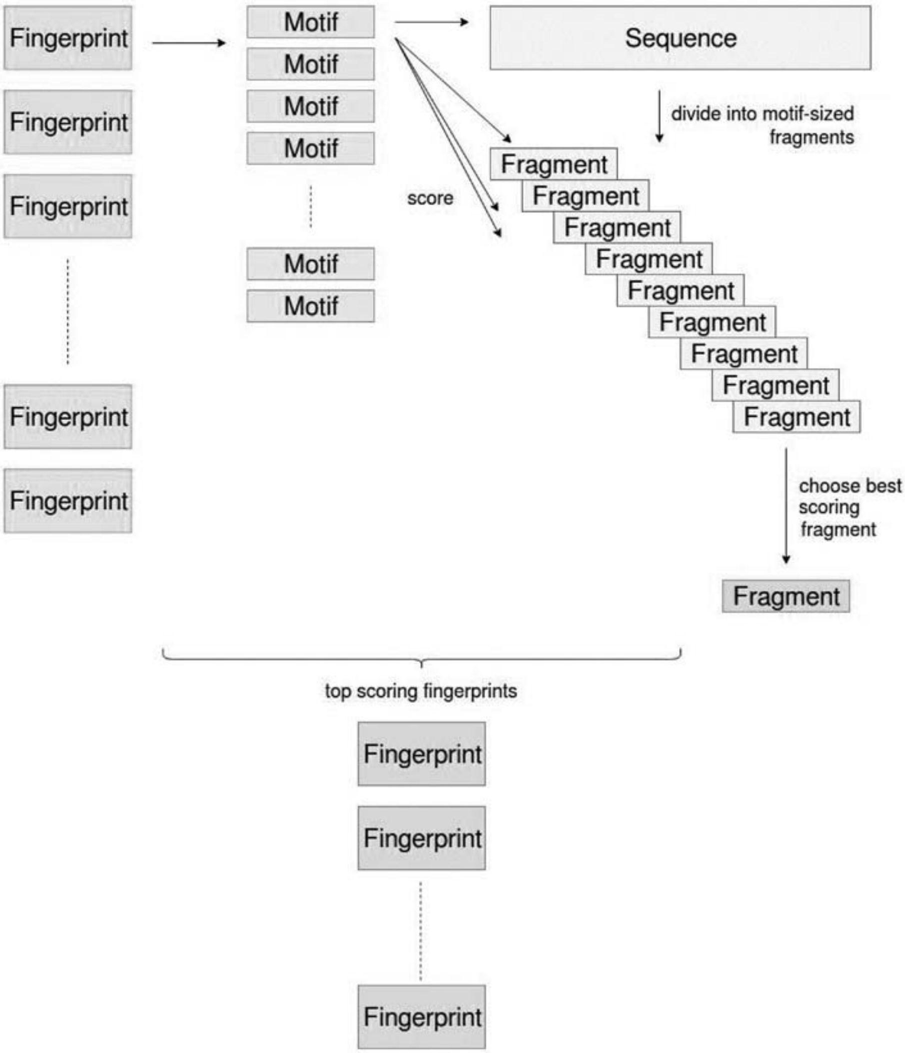


Figure 3 - Algorithm for scanning an unknown protein sequence against PRINTS

3 Models for object oriented design

The implementation of the protein fingerprint search algorithm will use an object-oriented model, based on the concept of objects containing data in the form of fields (attributes). Objects have functions known as methods that can access and modify the data of the object that they are associated with. The analysis of the problem requires it to be divided into separate components. Each such component is considered as an individual object or a set of objects between which communication is transmitted in the form of messages. There are two main types of classes used. The first type represents data from the relational PRINTS database via object-relational mapping (ORM). The second type of classes is abstract and represents the components of the algorithm as well and the obtained results. Below is an abridged explanation of the main classes used in the program:

3.1. Fingerprint, Motif, Sequence

These three classes contain the data retrieved from a database record of the corresponding table in PRINTS. This approach of using ORM allows associating of various methods with the objects, such as calculating the score of a fingerprint against a protein sequence or generating a histogram of amino acids in every position in a motif.

The **Fingerprint** objects are initialised with fingerprint information such as title, annotation, creation_date, update_date and relationship with motifs. It also coordinates message passing from the motif objects in order to identify matches.

Motif objects are initialised with motif information such as title, code, relationship with a fingerprint and sequences. Later, motif scores are passed back to the fingeprints.

Sequence objects are initialised with sequence data such as sequence, pcode and relationship with a motif. The main purpose of sequences is to participate in the creation of motif's frequency table and profile.

3.2. Substitution Matrix

It is scientifically proven that certain amino acids can be substituted with others in a particular way[4]. This information is stored in substitution matrices that describe the rate at which one character in a sequence changes to another over time.[6] Rows and columns describe amino acids and cells contain scores that describe the substitution of the two amino acids. Substitution matrices are used in the computing of motif profiles in the protein fingerprint search algorithm. They are stored in YAML files - a human-readable data serialization language, commonly used for configuration files. The **SubstitutionMatrix** class performs the loading of a particular matrix by a given type and retrieving the score for a given pair of amino acids.

3.3. Frequency Table

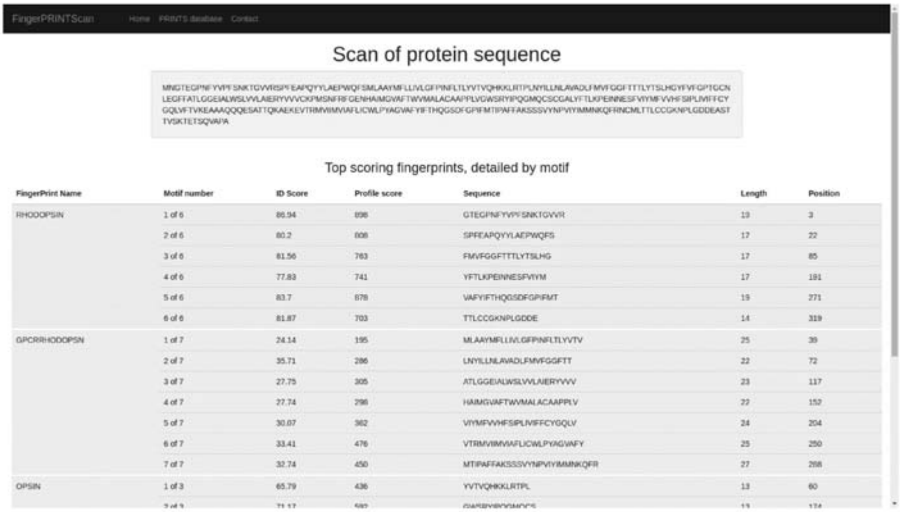
In order to analyze a certain motif, an instance of **FrequencyTable** class is created. It implemets the generating of a histogram (frequency table) of the amino acids on each position in the motif[5]. Another function of this class is the computing of a profile which utilizes both the histogram and the chosen substitution matrix. FrequencyTable also performs the scoring of a motif against the uncharacterized protein sequence and thus contains an important part of the search algorithm - breaking the query sequence into motif-sized fragments and returning the best scoring ones.

3.4. Match, Motif Score and FingerprintScore

To achieve a diagnostic identification, three types of measures are introduced in order to identify a "match" with a fingerprint - amino acid match, motif match and fingerprint match[8]. The amino acid match shows whether the amino acid on position X matches any of the amino acids in the frequency table on that position. If this condintion is not met, the score is 0, otherwise the score is equal to the number of time the amino acid is present in the frequency table. The score is summed for every position in the frequency table and motif score is returned. The highest score for every motif represents the fingerprint score. The following three classes implement the evaluation of the results from the search algorithm. Instances of **Match** class contain information about the profile, identity score, statistical significance, the fragment of the query sequence being analyzed and the position in that fragment. Every motif has a **MotifScore** object that builds a list of Match instances and provides statistics such as top match or average score of the matches. **FingeprintScore** instances represent the score of a given fingerprint, providing the number of motifs that score, i.e. meet a certain criteria, and the type of the match - full or partial. Full matches are defined as all motifs in a fingerprint consistently scoring high. Partial matches are matches with less than all of the motifs in a fingerprint[8].

4 Web application

The algorithm is implemented in **Python**, a high-level programming language that is widely used in science and in particular in bioinformatics because of its whitespace indentation and code readability. The web application is created using **Flask**, a lightweight BSD licensed framework for Python. Flask is a micro framework because it does not require additional libraries, however it is highly extendable and various components can be added. The web application has a modern and responsive look, achieved with Bootstrap - a popular open source front-end web framework for designing websites. The code of the program is publicly available[11] which will make the future development and maintenance of the projects easier.



Scan of protein sequence

MNGTEGPNFYVPFSNKTGVVRSPFEAPQYYLAEPWQFSMLAAYMFLLIVLGFPINFLTLYVTVQHKKLRTPLNYILLNLAVADLFMVFGGFTTTLYTSLHGYFVFGPTGCN
LEGFFATLGGEIALWSLVVLAIERYVVVCKPMSNFRFGENHAIMGVAFTWVMALACAAPPLVGWSRYIPQGMQCSCGALYFTLKPEINNESFVIYMFVVHFSIPLIVIFFCY
GQLVFTVKEAAAQQQESATTQKAEKEVTRMVIIMVIAFLICWLPYAGVAFYIFTHQGSDFGPIFMTIPAFFAKSSSVYNPVIYIMMNKQFRNCMLTTLCCGKNPLGDDEAST
TVSKTETSQVAPA

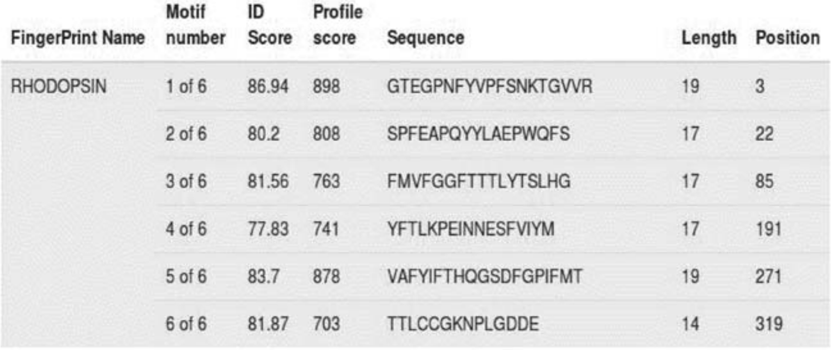
Top scoring fingerprints, detailed by motif

FingerPrint Name	Motif number	ID Score	Profile score	Sequence	Length	Position
RHODOPSIN	1 of 6	86.94	898	GTEGPNFYVPFSNKTGVVR	19	3
	2 of 6	80.2	808	SPFEAPQYYLAEPWQFS	17	22
	3 of 6	81.56	763	FMVFGGFTTTLYTSLHG	17	85
	4 of 6	77.83	741	YFTLKPEINNESFVIYM	17	191
	5 of 6	83.7	878	VAFYIFTHQGSDFGPIFMT	19	271
	6 of 6	81.87	703	TTLCCGKNPLGDDE	14	319
GPCRRHODOPSN	1 of 7	24.14	195	MLAAYMFLLIVLGFPINFLTLYVTV	25	39
	2 of 7	35.71	286	LNYILLNLAVADLFMVFGGFTT	22	72
	3 of 7	27.75	305	ATLGGEIALWSLVVLAIERYVVV	23	117
	4 of 7	27.74	298	HAIMGVAFTWVMALACAAPPLV	22	152
	5 of 7	30.07	362	VIYMFVVHFSIPLIVIFFCYGQLV	24	204
	6 of 7	33.41	476	VTRMVIIMVIAFLICWLPYAGVAFY	25	250
	7 of 7	32.74	450	MTIPAFFAKSSSVYNPVIYIMMNKQFR	27	268
OPSIN	1 of 3	65.79	436	YVTVQHKKLRTPL	13	60
	2 of 3	71.17	592	GWSRYIPQGMQCS	13	174

Figure 4 - Screenshot of the web application prototype

5 Conclusion

The program that is generated for this project achieves almost all defined golas. Testing was performed and the program successfully identifies the RHODOPSIN fingerprint as the best match for the amino acid sequence of OPSD_SHEEP. Moreover, it constructs a detailed list with scores for every motif present in the fingerprint (Figure 5). However, the calculation of statistical significance, presented as p-values, needs to be improved in order to provide more precise values.



FingerPrint Name	Motif number	ID Score	Profile score	Sequence	Length	Position
RHODOPSIN	1 of 6	86.94	898	GTEGPNFYVPFSNKTGVVR	19	3
	2 of 6	80.2	808	SPFEAPQYYLAEPWQFS	17	22
	3 of 6	81.56	763	FMVFGGFTTTLYTSLHG	17	85
	4 of 6	77.83	741	YFTLKPEINNESFVIYM	17	191
	5 of 6	83.7	878	VAFYIFTHQGSDFGPIFMT	19	271
	6 of 6	81.87	703	TTLCCGKNPLGDDE	14	319

Figure 5 - Detailed results for RHODOPSIN fingerprint, identified as top match for OPSD_ SHEEP sequence

This project can be extended with additional functionalities such as integration with UniProt:Swiss-Prot and UniProt:TrEMBL. This will allow users to submit sequence IDs instead of raw sequence strings as in the moment. Another feature that can be implemented is integration with other databases in order to provide PRINTS related information and classification to other tools.

6 Acknowledgements

The presented work has been funded by the Bulgarian NSF within the Sofia University SRF “Knowledge discovery in complex genome annotation analysis“ project, Contract No. 80-10-231 of 09 May 2017.

6 References

1. Teresa K. Attwood, Alain Coletta, Gareth Muirhead, Athanasia Pavlopoulou, Peter B. Philippou, Ivan Popov, Carlos Roma´-Mateo, Athina Theodosiou, Alex L. Mitchell (2012) *The PRINTS database: a fine-grained protein sequence annotation and analysis resource—its status in 2012.* Database (Oxford), Vol. 2012, Article ID bas019
2. T. K. Attwood, M. E. Beck, A. J. Bleasby and D. J. Parry-Smith (1994) *PRINTS - a database of protein motif fingerprints*. 3590 -3596 Nucleic Acids Research, 1994, Vol. 22, No. 17
3. Amos Bairoch, Rolf Apweiler (1999) *The SWISS-PROT protein sequence data bank and its supplement TrEMBL in 1999.* Nucleic Acids Research, 1999, Vol. 27, No. 1
4. Steven Henikoff, Jorja G. Henikoff (1992) *Amino acid substitution matrices from protein blocks*. Proc. Natl. Acad. Sci. USA Vol. 89, pp. 10915-10919, November 1992 Biochemistry
5. Paul G. Higgs and Teresa K. Attwood, *Bioinformatics and Molecular Evolution* (2005)
6. David W. Mount, *Bioinformatics: Sequence And Genome Analysis* (2001)
7. Philip Scordis, Darren R. Flower, Teresa K. Attwood (1999) *FingerPRINTScan: intelligent searching of the PRINTS motif database*. Bioinformatics, vol. 15 no. 10 1999, pages 799–806
8. Philip Scordis, *Diagnostic Identification of amino acid sequences Using the method of FingerPrinting*
9. Анатолий Димитров (2015) *Структуриране на биоинформатична релационна база данни PRINTS и съпричастни функционални уеб приложения*
10. DbBrowser - *Protein Families* http://www.bioinf.man.ac.uk/dbbrowser/ember/prototype/CHAPTER03/INFORMATION.shtml
11. Source code of the program https://bitbucket.org/mbdimitrova/fingerprintscan

Pattern Classification of Alternative Splicing in Gene Expression

Ognyan Kulev

Faculty of Mathematics and Informatics, Sofia University “St. Kliment Ohridski”,
5 James Bourchier blvd., Sofia 1164, Bulgaria
okulev@fmi.uni-sofia.bg

Abstract. Rapid development of sequencing technologies in molecular studies as well as in bioinformatics led to constantly changing annotations to genome data. What is lacking, are proper *in silico* evaluation methods for assessing the dynamics of complexity and comprehensiveness of these genome annotations. The complexity of gene models comprise variety of layers in genome annotations bearing the comprehensiveness of the annotation model which is aimed to be the most abundant and useful in practical biomedical studies. This paper presents current computational methods for classification of alternative splicing in gene models and how such classification could be used in assessing complexity and comprehensiveness of the applied models. Some open problems are discussed and a novel method for alternative splicing modelling and annotation is presented.

Keywords: genome complexity, annotation comprehensiveness, gene expression studies, alternative splicing.

1 Introduction

The biological process of gene expression follows the central dogma of molecular biology that DNA is transcribed into RNA and RNA is translated into proteins. Since RNA and proteins are the building blocks of all living cells, studying this process is essential for deepening our understanding of live. In higher organisms, these RNA and proteins have much wider variety than the genes from which they are produced. This is all thanks to gene expression and splicing events that allows a gene to produce multiple RNA products by splicing the DNA differently. Gene models represent alternative splicing as series of gene regions (exons) that are transcribed into one mRNA (transcript). One gene region can produce several alternative transcripts.

V. Dimitrov, V. Georgiev (Editors): ISGT’2017. ISSN 1314-4855
Proceedings of the 11th International Conference on
Information Systems and Grid Technologies, Sofia, 29.-30. Sep. 2017.

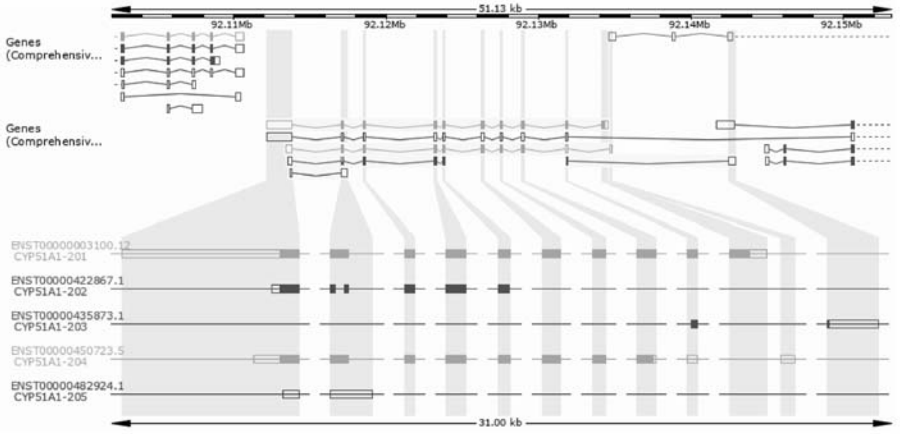


Fig. 1. Screenshot from Ensembl website that shows alternative transcripts of a gene.

In fact, gene models represent only part of the structural annotation of genomes but this part is very important from a practical point of view. Organisms have different cells, each in different development stage and living in different environmental conditions. Measuring gene product expression levels is a fundamental bioinformatics and genomics challenge for medical and research purposes. There are many tools doing such measuring and they all depend on gene model to map the expressed transcripts to the alternative transcript [1]. Changes in gene model can lead to opposite medical or scientific results and speculations.

The two problems that this paper presents and discusses are *in silico* evaluating the comprehensiveness and complexity of gene model. Our knowledge of genomes is always expanding and deepening but it is valuable to have assessment how much we know and if this knowledge is comprehensive enough to rely on. Such assessment is always relative to the current limitations of available technology, methods and data. When limitations change, our assessment of comprehensiveness may differ very much. For our purposes, it is needed that comprehensiveness is evaluated is evaluated only within genome reference and bioinformatics tools. Ultimately, the purpose of comprehensiveness metrics is to show how low is our expectation that gene model will change within some reference limitations.

Complexity is related to the complexity of the whole biological machinery involved in gene expression. Biological regulation of gene expression is represented as complex regulatory networks often involves many factors for single alternative transcripts [2]. Building regulatory networks knowledge base is very expensive and time-consuming process. For this reason, only small number of model organisms have extensive to a certain level regulatory networks. For the rest of organisms, we can only guess and infer how complex are their regulatory networks. In contrast, gene models are basic annotation products for any genome and usually they are starting point for studying new organisms. Choosing which specific alternative transcript to express is only small part of regulation, and it can represent the complexity of regulatory networks.

Different research groups study the same organism using different methods and references. The ability to objectively compare their gene model is valuable for making decision which specific gene model to use [3].

The major objective of the study is to assess the gene model complexity. The inference based on this assessment is related to the evolutionary complexity of the organism. The second objective of the study is development of a comprehensiveness metrics providing how we are close to the targeted biologically real gene model.

2 *In silico* modelling of alternative splicing

To assess comprehensiveness and complexity of a gene model, alternative transcripts must be compared and the differences have to be represented. When comparing two alternative transcripts, there are matching parts and differing parts. These differing parts are classified into patterns of alternative splicing.

A state of the art method for representing patterns is described in [4]. When comparing two alternative transcripts, the above method divides exons into smaller regions that either completely overlap in the two alternative transcripts, or there is overlapping matching intron region in the other alternative transcript. This allows encoding of each alternative transcript as series of 0 and 1, where 1 stands for part of exon region, and 0 is region not part of exon (that is, part of intron). Each binary digit from one alternative transcript has related binary digit from the other compared alternative transcript for the same gene region.

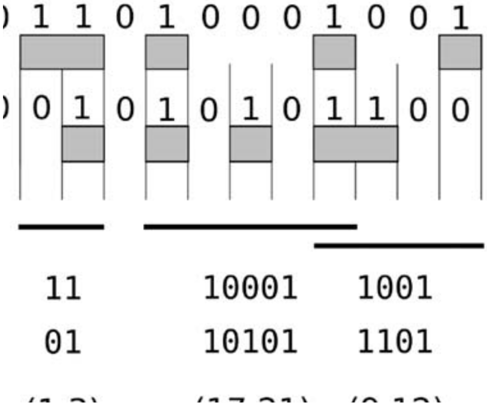


Fig. 2. Binary and numeric representation of patterns in alternative transcripts.

Using this method, a binary number represents an alternative transcripts compared to another alternative transcript. Pattern binary numbers use having both 1 in the two alternative transcripts as delimiter. Any pattern includes a binary place where there is 0 in one of the alternative transcripts, and 1 in the other alternative transcripts. From this binary digit we search the nearest delimiter on the left and nearest delimiter on the right. With this algorithm, any pattern can be detected and it unifies many alternative splicing events, independent of region sizes.

In the literature, there are seven recognized representative patterns. In reality, there are many more that can be constructed by combining the seven representative patterns, or by increasing and shuffling the participating exons. Running this method on model organism gene model generates thousands of patterns. This makes the described method suboptimal for true classification of the variety of alternative splicing patterns.

Another problem with this method is that works only on pairs of alternative transcripts but often gene models include more alternative transcripts per gene. To assess complexity, all alternative transcripts of gene should be taken into account.

In this study is developed a pipeline which uses the already existing implementation of this method, and we applied it to all human gene models found in Ensembl since version 43 (released ten years ago in 2007). The number of events for each pattern comprise pattern profile and it was used to study the dynamics of pattern changes in gene model releases. When conditions like genome reference assembly do not change, pattern profile smoothly raises the number of discovered patterns. But changes in genome reference assembly often abruptly change pattern profile. We concluded that comprehensiveness assessment needs to be limited only to specific conditions. When conditions change, there should be reset of comprehensiveness measurement.

3 Novel method in development

To solve the problems of [4] method, this paper proposes a novel representation of alternative transcripts. The method also splits exons into uninterrupted regions that are represented as 0 and 1. These uninterruptible gene regions we call “exon units” and they are not between two alternative transcripts but amongst all alternative transcripts of a gene. All alternative transcripts of a gene can be visually represented by a character grid where each row is alternative transcript and each column is exon unit region. Characters in cells are: “*” for part of exon in alternative transcript, “ . ” for part of intron in alternative transcript, and “ ” for region outside alternative transcript (before or after the alternative transcript itself).

```
ENST00000003100 ****.***..*.*.**.*.*.*.**.*
ENST00000450723  ***.***..*.*.**.*.*.*.**...*
ENST00000422867   **.*.*..*.*.*
ENST00000482924    *.****
ENST00000435873                         *.....*
```

Fig. 3. Visual text represenation of transcripts.

Now we can also compare pairs of alternative transcript and get binary representation of the comparison. To completely match this binary representation to [4] method, it is just needed all regions with the same binary digit X that correspond to repeating binary digit Y in the other alternative transcript to be replaced with single X and single Y in the two alternative transcripts respectively. Visually, patterns can be represented by enclosing patterns in parenthesis. Brackets are used when the pattern boundary is before the beginning of after the end of alternative transcript.

```
ENST00000003100 [*)***.***..*.*.**.*.*.*.**(.*    ]
ENST00000450723 [ )***.***..*.*.**.*.*.*.**(...*  ]

ENST00000003100 [**)**.*(*)*..*.*.*(*.*.*.*.**.*    ]
ENST00000422867 [  )**.*(.)*..*.*.*(                ]

ENST00000003100 [***)*.***(..*.*.**.*.*.*.**.*    ]
ENST00000482924 [   )*.***(*                      ]

ENST00000003100 [****.***..*.*.**.*.*.*.*)*(.*    ]
ENST00000435873 [                        )*(.....*]
```

Fig. 4. Visual text representation of pairwise pattern comparison between transcripts.

The places of the parenthesis have biological significance. They are points where biological splicing machinery switches between alternative splicing depending on biological regulation. Thus, having all these alternative splicing points is important for assessing the complexity of gene model. These points can be visually represented as “|” characters.

```
[*|*|*|*.*|*|*|.|.*.*.*|*.*.*.*.*|*|.*    ]
```

Fig. 5. Visual text representation of gene exon units with alternative splicing points.

These alternative splicing points belong to the structural annotation of the gene. They can be added as new type of structure annotation records. To be useful for reconstructing patterns, they need to include additional attributes to the annotation. Each splicing point includes set of pairs of alternative transcripts that have different splicing beginning at this point. In other words, when comparing two alternative transcripts, there would be opening parenthesis at this place in the visual representation of alternative splicing patterns. Similarly, there is a separate set of pairs of alternative transcripts where the splicing switches from being different to being the same (closing parenthesis in visual representation). Such sets of pairs are undirected graphs where nodes are alternative transcripts and edges are pair of alternative transcripts. All graphs in such structural annotation for a gene share the same nodes (alternative transcripts) and only differ in their edges.

The sequence of graphs matches the biological sequence of splicing decisions made when DNA is transcribed into RNA and that indirectly reflect the complexity of regulation and the organism as a whole. The method is still in development and the desired model for structural representation of alternative splicing patterns is not yet developed.

4 Conclusion and future development

Assessing the complexity and comprehensiveness of gene models is unexplored area that will bring practical benefits to biomedical research. The current state of the art method for classifying alternative splicing events is presented and its shortcomings are discussed. A new method for modelling alternative transcripts of gene is being developed and its current state of development is explained. In future work, the method will be evaluated and validated for assessing complexity and comprehensiveness of mammal genomes.

Acknowledgements. The presented work has been funded by the Bulgarian NSF within the Sofia University SRF "Knowledge discovery in complex genome annotation analysis" project, Contract No. 80-10-231 of 09 May 2017.

References

1. Oshlack, A., Robinson, M. D., Young, M. D. (2010). From RNA-seq reads to differential expression results. *Genome biology*, *11*(12), 220.
2. Hasty, J., McMillen, D., Isaacs, F., Collins, J. J. (2001). Computational studies of gene regulatory networks: in numero molecular biology. *Nature Reviews Genetics*, *2*(4), 268.
3. Zhao, S., Zhang, B. (2015). A comprehensive evaluation of ensembl, RefSeq, and UCSC annotations in the context of RNA-seq read mapping and gene quantification. *BMC genomics*, *16*(1), 97.
4. Nagasaki, H., Arita, M., Nishizawa, T., Suwa, M., Gotoh, O. (2006). Automated classification of alternative splicing and transcriptional initiation and construction of visual database of classified patterns. *Bioinformatics*, *22*(10), 1211-1216.

The Methodology behind Applying Adapted QSAR as a Cost-Effective Method for in Silico Biotechnology Experiments

Valeriya Simeonova, Dimitar Vassilev

Faculty of Mathematics and Informatics, Sofia University “St.Kliment Ohridsky”
Sofia, Bulgaria
{simeonova, dimitar.vassilev}@fmi.uni-sofia.bg

Abstract. One of the mager problems in biotechnology experiments is still the cost and time, to obtain needed results. Here is described the methodology behind applying an adapted QSAR as a cost and time effective method of identifying the most theoreticaly possible combinations of nutrient media, that produce high results in biotechnological point point of view. In other words, the aim is to find a cost-effective and time-effective method of identifying nutrient media producing identical plants with maximum performance in terms of the bioactive substances contained in them. All produced in silico nutrition media are based on the ranges of phytonutrient hormones in biotechnological experiments. A big part of the methodology is data preparation, as they come most of the time in custom format and views, produced by the biotechnology researcher. In most cases, biotechnology experiments are firstly done and after that is seeking a mechanism to analyze them. In vary rare cases experiments are carefully planned with the data analisyst. The obtained results can be used as: a theoretical guideline for determining the optimal nutrient media and combinations; to study other medicinal plants in order to establish effective biotechnological schemes for growth and rooting that are also cost-effective; using ANN, taking into account the species and the ecotype.1

Keywords: Methodology, Artificial Neural Networks, adapted QSAR, in vitro experiments

Abbreviations: AC ratio – cytokinin /auxin ratio; AI – Artificial Intelligence; ANN – Artificial Neural Network,; BAP – N6-benzylaminopurine; CM – Culture Meduim/a; 2,4-D – 2,4- dichlorophenoxyacetic acid; GA3 – gibberellic acid; IAA – Indolyl-3-acetic acid; 2-iP – 6- (y,y-dimethylallyl amino) purine; IBA – Indole 3-butyric acid; IPM – Initial Plant Medium/a; MC – Media Combination; NAA - a- naphthyl acetic acid; TDZ – thidiazuron; QSAR - Quantity Structure-Activity Relationship

V. Dimitrov, V. Georgiev (Editors): ISGT’2017. ISSN 1314-4855
Proceedings of the 11th International Conference on
Information Systems and Grid Technologies, Sofia, 29.-30. Sep. 2017.

1. Introduction

In recent years, the in vitro culture method has been used as one of the advanced biotechnology systems to produce a large number of identical plants for a short period of time that are free of pathogens. Such requirements are posed by the needs and demand of industries such as horticulture, agriculture and forestry. This is particularly useful for:1) species with wide or massive use; 2) with slow and difficult cultivation under natural conditions; 3) medicinal plants with intensive use, which leads to the extinction of some species, the limitation of natural populations and biodiversity.

Therefore, it is important to find a cost-effective and time-effective way of identifying nutrient media producing identical plants with maximum performance in terms of the bioactive substances contained in them.

The concept of "cost-effective" is linked on several ways:

- Artificial tests with a high range of materials (often limited source in real world)
- Artificial tests with a high range of consumables provide:
    - Once: identifying the low cost consumables, producing *identical plants with maximum performance in terms of the bioactive substances contained in them.*
    - Second: a chance to make in vitro experiments only with the best from *Once*

The concept of "time-effective" is linked to:

- Artificial tets with a high range of consumables need in way less time than in-vitro experiment
- The chance to make in vitro experiment only with the best results from *Once* shorten the time to discover the best performance in biotechnologycal aspect.

This allows biotech studies to be conducted in "depth", ie. massive investigation of effect in small differences in concentrations.

The Quantitative Model of Structure-Activity Dependency (QSAR) for analysis is interdisciplinary and uses knowledge in pharmacology, molecular biology, organic and quantum chemistry, analytical methods for structure analysis, mathematical and engineering methods, statistics, informatics, etc. [1]. QSAR is a method based on the hypothesis that similar biological activity is determined by common structural characteristics. The aim is to quantify the relationship between the chemical components (in our case, the concentrations of phyto-regulators and other components of the nutrient medium) and the biological activity they provoke. Analyzes are often graphic and depending on the dimensionality of the data is subject to different two-dimensional or three-dimensional graphic models, and any combination thereof, as projected data are multidimensional. Three-dimensional QSARs are methods used to detect the quantitative link between the spatial structure of the chemical component and its activity. [2].

The need to adapt QSAR is provoked by several reasons:

- Originally is used in drug discovering
- In drug discovering the 3D model represents the molecular structure
- With appropriate adaptations could be used for describing any relationship
- In the case of in vitro experiments for bioactive substance:

o it describes the relation: process success ~ medium & time period & material

o the 3D model represents cluster formations with combinations of media that assures in less time highly production of in vitro plants with minimum cost. We called 3D because of the three dimensions of the analysis.

o the "spatial structure of the chemical component," is replaced with "the multidimensional structure of the nutrient medium as determined by the different phytohormon concentrations and other quantifiable environmental and experimental conditions" or with the time intervals in which explants are treated with various decontaminating chemicals.

2. Material and methods and software

The biological experiments:

The data used for bioinformatic analysis are from biological experiments conducted in 2 to 3 iterations. The results obtained are summarized and described in detail [6]– [9].

Bioinformatics experiment: Each biological process (germination, propagation, rooting and two tables for decontamination) is represented by different ANN. This means that we have five data sets and five architectures, by one for each data set. In this manner, we have five different training processes. All results are represented in a single table.

Microsoft Office Excel was used for data preparation, query preparation and all graphic representations for the analysis.

EasyNN Plus [3] is a software, which works like a frame to construct ANN. The application uses back-propagation training algorithm. The software works with different type of files for input/output like: excel [4]

workbooks, text files and .CSV files. It is implemented real-time graphic generation and visualisation, including for estimating errors. It is possible to import an excel file with queries after training process is done. All of the results are exported and further analysed. The verification mechanism includes statistics and it is based on the average error value less or equal to 0.0001.

The forecasting stages are developed in the context of the process controlling of tissue and roots formation because of their role in in-vitro reconstruction [5].

Artificial Neural Networks (ANN): Neural Networks (NN) is a subclass of Artificial Intelligence (AI) methods. It is known that AI is the first step of the QSAR methodology, respectively Adapted QSAR. ANN are chosen because of their strong impact in forecasting and classifying data. The method is applied on data with high percentage of uncertainty and imperfection.

The methodoly of the analysis is given on the Figure.1

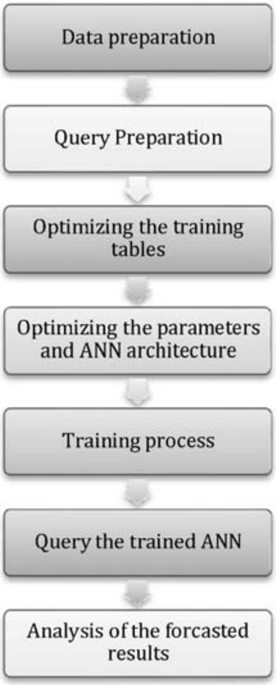

Data preparation
Query Preparation
Optimizing the training tables
Optimizing the parameters and ANN architecture
Training process
Query the trained ANN
Analysis of the forcasted results

Figure 1: The methodology at a glance

3. How does look the initial data?

- The initial data comes aggregated, described with standard deviation
- We have 4 processes: cleaning, germination, propagation, rooting
- Material:
  - seeds - fresh and old
  - Explants – buds, adventive buds, root buds, stem segment, root segment

- Price lists for:
  - cleaners
  - medium nutrients and phytohormones
- Treatment days

4. How to prepare the data?

- Desaggregarion process is needed, because often data comes already aggregated. For this process, all is needed:
  - generators for normal distributed data
  - the mean
  - standard deviation
  - the results
- Identifying the initial input neurons for the ANN: all supporting information about current process is used
- Generate the calculated input neurons: supports forcasting and conclusions:
  - Medium_value = $\sum_{\text{��}=1}^{n} \binom{\text{��}}{k} \mathrm{p}_k \text{��}_k$, where: pk is the price and q_k is the quantity of the k-th phytohormone
  - Medium_coefficient: calculates the auxin/cytokinin ratio, based on phytohormone classification
  - Medium_type: based on the medium_coefficient there are:
    - Type 0 with medium_coefficient in [1;24] – stimulates growth and development of the plant
    - Type 1 with medium_coefficient in [0;0.5] – stimulates root formation
    - Type 2 with medium_coefficient in (0.5;1) – stimulates as well as growth, development and root formation and caluses
  - G.Price.Days = GD*medium_value/10, where GD is germination days. This field compensate the lack of data due to the germination days was excluded at the stage of optimizing the training table
  - MC_Price = G.Price.Days + Rooting price
  - + a lot of other calculation fields supporting those we finally used for training such as: the phytohormone classification as auxin ot cytokinin, price list, price calculation by unit, vlookup tables
- Training tables constructions: for each biotechnological process, there is a separate training table, as for the process of cleaning they are two – one for the seeds and one for the explants.

Table 1: The full training table structure, describing all biotechnological processes

Column name	Seeds Cleaning	Explant Cleaning	Germi-nation	Propa-gation	Rooting	Data Type	Field Type
Plant Material	x	I	x	x	x	Text	E
Ache 5%	I	I	x	x	x	Integer	E
Bleach 2,2%	I	I	x	x	x	Integer	E
Ethyl Alcochol (C2H5OH) 70	I	I	x	x	x	Integer	E
HgCl2 (0,2%)	I	I	x	x	x	Integer	E
Peroxyd (2,2) (3%)	I	I	x	x	x	Integer	E
Seeds	O	x	x	x	x	Integer	E
Period	I	x	x	x	x	Integer	E
Cleaned	O	O	x	x	x	Real	E
Development	O	O	x	x	x	Real	E
GA3	x	x	I	E	I	Real	E
Sucrose	x	x	I	E	I	Real	E
Agar	x	x	I	E	I	Real	E
Days	x	x	I	x	x	Integer	E
VegetationPersentage	x	x	O	x	x	Real	E
High	x	x	O	x	x	Real	E
Explant_Type	x	x	x	I	x	Text	E
Zeatin	x	x	x	I	x	Real	E
BAP	x	x	x	I	x	Real	E
Kinetin	x	x	x	I	x	Real	E
2-iP	x	x	x	I	x	Real	E
IOK	x	x	x	I	I	Real	E
ANO	x	x	x	I	I	Real	E
Glutamine	x	x	x	I	x	Real	E
Cazeine	x	x	x	I	x	Real	E
2,4 D	x	x	x	E	x	Real	E
TDZ	x	x	x	E	x	Real	E
IMK	x	x	x	E	I	Real	E
Medium_Value	x	x	x	I	x	Real	C
Medium_koeficient	x	x	x	I	x	Real	C
Medium_Type	x	x	x	I	x	Integer	C
Medium_Type_test	x	x	x	I	x	Integer	C
New_Shoots	x	x	x	O	x	Real	E
shoots high	x	x	x	O	x	Real	E
G.GA3	x	x	x	x	I	Real	E
G.Sucrose	x	x	x	x	E	Real	E
G.Agar	x	x	x	x	E	Real	E
G.Days	x	x	x	x	E	Integer	E
Rooting	x	x	x	x	O	Real	E
G.Price.Days	x	x	x	x	I	Real	C
MC_Price	x	x	x	x	I	Real	C
Pro_Index	x	x	x	x	E	Real	E

Table legend:

- X: the field is not included
- I: the field is included as input neuron
- O: the field is included as output neuron
- E: the field is excluded due to lack of data

Field type:

- E: Emperical
- C: Calculated

5. Query preparation

It is well known that if need to query a trained table, it is need that queries include the information for the input neurons. That's why query tables were easy to create as different combinations of phytohormon quantities with all supporting calculated fields, included as input neurons. It important to say that if biotechnology experiment explores the upper and down values of quantities, the training and queries will give more accurate results.

6. Optimizing the training tables

Easy NN Plus provide optimizations on the training tables mainly in two directions: 1) Filters columns with same values; 2) Provides analysis of importance and sensitivity. Regarding the data in this case, this stage includes:

- For each neural network, there is analysis of importance
- For each neural network, there is analysis of sensitivity

- A joint analysis is made in form of scatter plot:

o X: importance
o Y: Sensitivity
o Input neurons are classified in 4 categories: we prefer high importance + high sensitivity, but doesn't exclude those with low sensitivity

The cathegories are:

Quadrant 1:
- the most undesired situation
- Result is slightly or even not reflected by

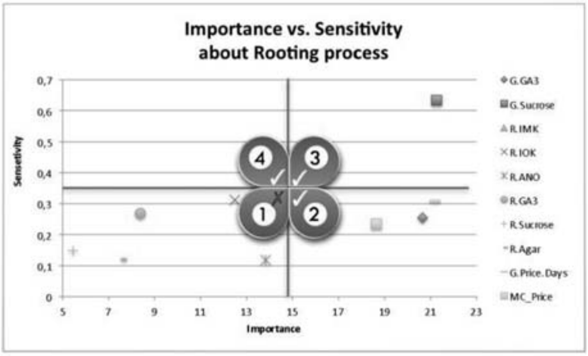


Figure 2: Example of a joint analysis of Importance and Sensetivity

- There is a need of huge changes to make a difference in the result

Quadrant 2:

- Better than (1 and 4) and worse than 3
- Result is reflected by
- There is a need of huge changes to make a difference in the result

Quadrant 3:

- the most desired situation
- Result is totally reflected by
- Little changes make a huge difference in the result

Quadrant 4:

- Better than 1 and worse than 3
- Result is slightly reflected by
- Little changes make a huge difference in the result

7. Optimizing the parameters and ann architecture

Easy NN Plus gives the opportunity to set manually all the initial parametres of the

training process, architecture of the hidden layers, stop conditions and error rates. Nevertheless, there are several optimizing controls, so the platform itself can find the most appropriate as:

- Parameters:
  - Learning rate
  - Momentum
- Architecture
  - The number of hidden layers
  - The number of neurons in hidden layers

Table 2: The optimized parameters and architechture structure of AINN

Parameters of AINN	S. C	Expl.C	G	P	R
Learning cycles	1335	219372	114454	1528	44673
Training error	0,0001	0,00009	0,0001	0,000098	0,000007
Input columns	6	6	4	13	9
Output columns	3	2	2	3	1
Serial columns	0	0	0	0	0
Excluded columns	0	0	0	6	4
Training example rows	12	19	24	15	27
Validating example rows	0	0	7	4	5
Querying example rows	200	512	357	1030	200
Excluded example rows	0	0	0	0	0
Duplicated example rows	0	0	0	0	0
Input nodes connected	7	6	4	13	10

Hidden layer 1 nodes	8	6	6	12	8
Hidden layer 2 nodes	4	6	6	6	6
Hidden layer 3 nodes	0	7	7	0	7
Output nodes	3	2	2	3	1
Serial input nodes	0	0	0	0	0
Serial output nodes	0	0	0	0	0
Learning rate	0,6	0,6	0,3902	0,4676	0,8859
Momentum	0,8	0,8	0,8709	0,8718	0,5472
Target error	0,0001	0,0001	0,0001	0,0001	0,0001
Validating error	N/A	N/A	0,078231	0,036244	0,006815
Validating example rows	N/A	N/A	100%	100%	100%
Adapting.	No	No	Yes	Yes	Yes

Each column (S.C., Expl.C, G, P, R) represents each one of the five ANNs. The maximum estimated error level is under 0.00001 in accordance with the predefined rules. There are two approaches that can be applied with the limited data collections:

a) Automated similiuation that multiplies similar training examples (this option is provided and used in EasyNN). Could lead to pretrenned results, but in the context of the research it is desired situation. As compensation to the limited data collection it is not provided the validation step of the trening procedure.

b) Using the standard deviation and the number of media, can be prodused as much examples as we need. This research is the next step. It should provide more accurate analysis.

8. Analysis on the forecasted results

As it is all about the methodology, one of the main questions is how many analyses do we have on the forecasted results? The answer is shown on the next graphic:

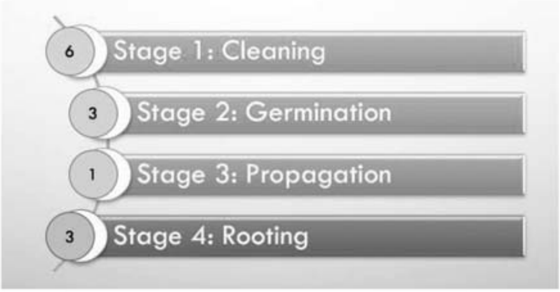


Figure 3: How many analyses do we have on the forecasted results?

STAGE 1: CLEANING Analysis (1.1): number of cleaning schemas that assures the best development results by type of seeds (old and fresh) Analysis (1.2): distribution of

1.1 by total cleaning time and choosing those with best development progress in short term cleaning schemas

Analysis (1.3): define the characteristics of the most time-effective cleaning schemas.

The same three analysis were provided for explants too.

STAGE 2: GERMINATION

Analysis (2.1): number of media for the best germination rate by culture period Analysis (2.2): distribute the best results from (2.1) by price range and filter the most cost-effective ones

Analysis (2.3): define the characteristics of the most cost-effective media

STAGE 3: PROPAGATION

Analysis (3.1): number of media for the best Propagation index, highest shoots' high

and lowest price range

Observed: the propagation index is proportional to the shoot's high

Remark: Could be done with radar graphics

STAGE 4: ROOTING

The rooting training table includes media parameters of germination process, following the biotechnological process, but also following the criteria for cost- effectiveness:

- germination_percentage >60%
- Germination_medium_price <1
- Cultivation days >=20

Analysis (4.1): number of media combinations (i.e. germination + rooting medium)

for the best rooting rate by cultivation days

Analysis (4.2): distribute the best results from (4.1) by price range and filter the most cost-effective ones

Analysis (4.3): define the characteristics of the most cost-effective media combinations

9. Results and disscussion

Results on real data shown that there are 43 theoretical combinations of media, describing the whole biotechnoly experiment, that promiss more than 80% success in the price range of [0-1,5] euro/liter.

Wet lab experiments are needed to verify theese results.

10. Conclusions

The obtained results can be used as a theoretical guideline for determining the optimal nutrient media and combinations thereof so that they are effective both from a biotechnological point of view and from an economic point of view. The approach can also be applied to the study of other medicinal plants in order to establish effective biotechnological schemes for growth and rooting that are also cost-effective. Possible perspectives are the creation of neural networks, taking into account the species and the ecotype, in order to build a more comprehensive system based on already available experimental data.

Acknowledgements

The presented work has been funded by the NSF project "GloBIG: A Model of Integration of Cloud Framework for Hybrid Massive Parallelism and its Application for Analysis and Automated Semantic Enhancement of Big Heterogeneous Data Collections" project, Contract ДН 02/9 of 17.12.2016

Bibliography

[1] J. Egelund, M. Skjøt, N. Geshi, P. Ulvskov, and B. L. Petersen, "A Complementary Bioinformatics Approach to Identify Potential Plant Cell Wall Glycosyltransferase- Encoding Genes," *Plant Physiol.*, vol. 136, no. 1, pp. 2609–2620, 2004.

[2] N. Nikolova, "Computer modeling of chemical structures. PhD Thesis," Technical University - Sofia, Sofia, Bulgaria, 2002.

[3] Neural Planner Software, "EasyNN Plus." 2012.

[4] Microsoft, "MS Excel." .

[5] Y. H. Su, Y. B. Liu, and X. S. Zhang, "Auxin-cytokinin interaction regulates meristem development," *Mol. Plant*, vol. 4, no. 4, pp. 616–625, 2011.

[6] K. Tasheva and G. Kosturkova, "Establishment of callus cultures of Rhodiola rosea Bulgarian ecotype," *Acta Hortic.*, vol. 955, pp. 129–135, 2012.

[7] K. Tasheva and G. Kosturkova, "The role of biotechnology for conservation and biologically active substances production of Rhodiola rosea: Endangered medicinal species," *Sci. World J.*, pp. 1–13, 2012.

[8] K. Tasheva and G. Kosturkova, "Rhodiola rosea L. in vitro cultures peculiarities," *Sci. Bull.*, pp. 103–111, 2010.

[9] K. Tasheva and G. Kosturkova, "Bulgarian golden root in vitro cultures for micropropagation and reintroduction," *Cent. Eur. J. Biol.*, vol. 5, no. 6, pp. 853–863, 2010.

Problem Space of the Monitoring Granularity in Cloud Systems

Vasil Georgiev, Daniel Simeonov, Hristo Hristov

Faculty of Mathematics and Informatics
University of Sofia «St Kliment Ohridski»
v.georgiev@fmi.uni-sofia.bg

Abstract. This paper presents a systematic scheme of the approaches to the cloud service monitoring systems. Monitoring is the core of overall performance management in the cloud systems. It determines the effectiveness of cloud control decisions and system actions as well as the performance cost measured as by the generated system overloads. The information precision of the current system state is orthogonal to monitoring system overload. We describe the linking category between these two concepts as monitoring granularity and present a systematization of possible approaches which are based on the concept of granularity.

1. Introduction

All the cloud systems are deployed in four possible use cases: private clouds, IAAS (i. e. virtual machines), PAAS (i. e. middleware components), and SAAS (i. e. application services) – Fig. 1. [11].

In all these four cases the key parameter of the service process is the efficiency of the deployed infrastructure. Another service parameters like availability, troubleshooting, service times, etc. QoS parameters are of less importance especially what concerns cloud servicing. Therefore, the resource-efficient control is essentially the most important process of cloud management. It is traditionally and logically decomposed into three processes - monitoring, planning and enacting (e. g. process and/or data transfer). Cloud virtualization allows one to combine effective planning of deployed cloud resources with the required quality of service to its users at the price of various system overload. The system overload is determined by the three processes – monitoring, planning and enacting. Monitoring can be characterized by the number of observed parameters and their spatial and temporal sampling. The planning algorithm (and hence the enacting) is dependent on the resource status and cloud service information.

If we should represent the architecture of those three control processes it will form pipeline or layered stack with the monitoring at its core or base. In this trinity monitoring is the decisive processes, as the other two are either direct (planning) or indirect (enacting) dependent on it. This dependence can be expressed as calls of services of the three processes: enacting aggregates planning methods which in turn aggregates monitoring methods. Thus the monitoring strategy is crucial for the entire cloud management process. Here we present a parameterization of this process by introducing the term of compound monitoring granularity. By this concept monitoring overload can be characterized by the number of observed parameters and their spatial and temporal sampling discretization. The planning process is a sharing or matchmaking of the arriving tasks or service calls between the deployed resources or service instances respectively, based on one or more optimization parameters. Typically, the purpose of planning is to achieve a higher quality of service, consisting of the shortest service time (Goal A) and/or the lowest error- and failures rates in the service system (Goal B). The usual approach to achieving Goal A is balancing the resource load in the distributed service system. In Section **2.**, we are dealing with the question of what numerical parameters measure the load and what - the resources. The usual approach to solving Goal B is applying of fault tolerance approaches typically being implemented by resource-consuming replication of the service processes. In other words, fault tolerance service means doubling the necessary resources. Here, under service quality, we will only look at optimizing for Goal A.

In order to analyze or evaluate the efficiency and the cost of the monitoring process in a distributed system one needs a classification scheme or set of axes on which to map various monitoring schemes. In **2.** we present a three-dimensional space for [quasi-]quantitative monitoring taxonomy. The values of each of the axes of this space are qualified in the scale “coarse” to “fine” in resemblance to a fundamental concept in the distributed systems’ performance which is the concept of granularity.

Granularity takes into account the communication and synchronization overhead between multiple processes or processing components. It is defined as the ratio of algorithmic computation time to communication and synchronization time, wherein, computation time is the time required to perform the computation of a task and communication time is the time required to exchange data between processors including necessary time for synchronization [6].

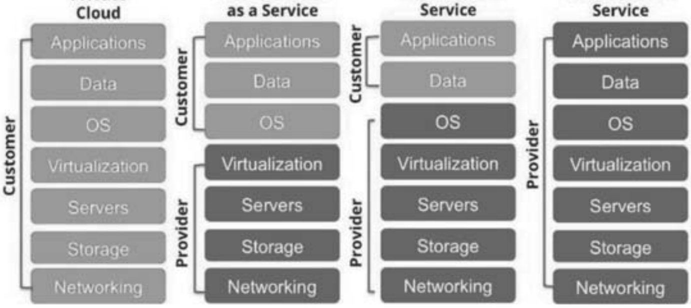


Figure 1. *Four use cases and seven layered components of the cloud infrastructure [11].*

The fine-grain concurrent processing which allows more precise load balancing between the processing cores or nodes. Analogously a complex and detailed monitoring process helps to control the cloud processing components in way to perform more efficiently. However such preciseness is at the price of bigger system overload. By “fine-grain” monitoring not only the system information procedures are more intrusive and resource consuming but also the control enacting decisions procedures are more complex and require more often intrusion into the general process of cloud orchestration. On the contrary “coarse-grain” monitoring keeps information just of the major performance parameters and at a lower sampling rate. As a result the system overload is low but the control procedures are just capable to keep services running without much care about resources’ efficiency. Consequently we have to trade between more efficient resource planning at higher system overload and less efficient one but with lesser overload. The evaluation of the system output is to be measured on the terms and scales of QoS (service time, etc.) but also on in terms of resources’ effectiveness – utilization, total number of customer tasks and services, sometimes even the electrical power consumption (e.g. the concept of green clouds).

Consequently, it is difficult if not impossible to predict or to synthesize a priory optimal level of the monitoring granularity. The first step in monitoring analysis and evaluation is to define a parameter which to compare and further to evaluate the complex system process of monitoring. We call this parameter granularity not only because of the cited resemblance to the parallel processing granularity but also because such a parameter cannot always be strictly quantified by a [scalar] figure. It might need a qualitative estimation on a scale like fine-grain ↔ coarse-grain granularity.

Furthermore the monitoring granularity is not a scalar or unidimensional concept. We present in 2. a **three-dimensional** monitoring granularity space comprising of sampling rates, state precisions (i.e. the number of distinguishable states) and the number of monitored resource parameters. Ideally or abstractly any of these three dimensions of the monitoring granularity is independent to the other two. Thus the concept of granularity is to be presented by a 3-ple estimative which is given in 2. In section 3. a case study illustrates application of the three-dimensional estimation of the monitoring granularity illustrate how our monitoring granularity scheme reflects the “real-world” system monitoring metrics. It is reasonable to speculate about the possibility to produce a single generalized parameter out of the monitoring granularity triple. We do not propose yet such a generalized scalar estimation but discuss this subject in the Conclusion.

2. Monitoring Granularity Scheme

The **monitoring process** is a registration and recording of the current values of one or more parameters (either numerical or qualitative) that characterize the state of service and/or resource over a certain "short" interval of time. Monitoring is therefore a time driven process. Obviously here the concept of "**granularity**" occurs as a product of three components:

❶ ***cardinality*** χ of the parameters' set (i.e. the number of the monitored system components);

❷ ***sampling rates*** $\boldsymbol{\sigma}$ of each of the monitored parameters, and

❸ ***precision*** $\boldsymbol{\pi}$ of the monitored parameters.

For referencing purposes we denote monitoring granularity and its components $\Gamma(\chi, \sigma, \pi)$. Furthermore, this multidimensional granularity can only be defined in the context of the service requirements, i.e. the load model parameters. For example if the load parameters change slowly with no burst task arrivals the sampling rate can be considered fine even for the longer sampling period. Thus monitoring granularity is dependent to the interarrival process of the services requests. Also we refer to this multidimensional monitoring parameter as granularity in resemblance to the load granularity. Coarse-grain tasks granularity usually is combined with longer interarrival periods and thus with shorter queues of waiting tasks. Vice versa fine-grain tasks are normally combined longer with queues of waiting tasks.

Under precision $\Gamma(\pi)$ we understand the number of the distinguishable states for a given performance parameter rather than merely its numerical precision. For example a popular monitoring parameter such as the resource utilization can be valued in percent which produces a scale of 100 distinguishable states of the monitored resource. We should consider this percentage scale of the utilization as the finest granularity: hardly there would be useful to measure utilization more precisely than this i.e. recording the fraction of the percentage. On the contrary: for resource control purposes a much coarser scale can be applicable and practical too e.g. comprising of four states – idle, underload, overload, bottleneck – which is an example of coarse monitoring granularity. Of course one can go a bit further in this direction defining just two resource states – e.g. underload and overload.

The planning process – on the contrary to the monitoring process – is event-driven one. The triggering event of it is the arrival of a new task or a new service call[1]. Of course sometimes so called burst arrivals or group calls occur. However the usual approach for the planning purposes is to execute matchmaking allocation

[1] This is so called arrival planning. There are departure planning strategies aimed to cure the situation of resources becoming [near-]idle by transferring partially executed tasks from non-idle resources. Such strategies are considered non practical and not applicable because they incur transferring of partially executed tasks at a heavy overload cost.

strategy task-by-task or call-by-call. Examples of such serial ordering of the concurrent or burst arrivals are countless. We'll just refer here to the ubiquitous Lamport's timestamps algorithm [7]. Therefore one has to transfer the time scale of the monitoring sampling to the event scale of planning. Such a scale needs a simple translation scheme comprising of several (and not much) distinguishable states of the monitored resource.

Looking back to monitoring process in distributed and cloud systems monitoring, different authors, under different assumptions, indicate two main thresholds for the use of a monitored component (such as its system queue length or its usage rate):

a) low threshold corresponding to idle and underload resource state and divides the status of the monitored component into idle and normal state. It is mainly applied to low-level monitoring (i.e. resources) and aims to prevent conditions where a service component is idle while there are other resources in the cloud with non-zero system queue or non-zero utilization. Typical values for this threshold are the zero or near zero values of the length of the resource system queue. According to some authors [5], most valuable monitoring information is to record only the zero state and few adjacent nearby states, in order to prevent the observed resource from falling idle.

b) high threshold corresponding to overload and busy resource state. The high threshold is applicable to both to low-level and high-level monitoring (i.e. to resources and services respectively). It is aimed to identify a component as an appropriate job donor or typically as inappropriate for targeting new incoming queries. A typical value for this threshold is, according to most authors, a 70% usage ratio, but some models [2] identify a deterioration of nonfunctional service parameters only at 80-85%. This threshold divides the status of the observed component as a normal state and overload.

c) a monitoring system of several thresholds. If we only apply a) or only b) we implement a 2-state monitoring system. If we apply a combination of both thresholds we get a system with three states – idle-to-underload, moderate, and overload. Of course, it is possible to add more thresholds and distinguishable states. In distributed system monitoring, space is formed by the number of distinguishable states of the monitored resource multiplied by the number of replicas of this resource. In cloud systems this number of resource replicas can be significant, e.g. datacenters and other public cloud systems.

In any case the utilization can be measured in the scale from 0 to 100%. If we measure it with the number of tasks in the system queue, we need to take into account the load granularity. For fine task granularity, the system queue length of several tasks can be interpreted as almost idleness of the resource, whereas in coarse task granularity the same number of tasks can be interpreted as a moderate load or even overload.

Consequently this first component of the monitoring granularity Γ(χ) can be implemented in two independent sets of parameters. One option is it to comprise of resource-oriented parameters i.e. utilization, idle time, capacity, etc. Such a set of parameters refers to a resource-oriented or low level of monitoring. Why this level should be "low" is justified by the cited diagram on Fig. 1. where it includes provider-oriented parameters. Another option is to do monitoring on the parameters of the higher customer-oriented level i.e. service performance parameters like service time and queue time.

The whole space of the monitoring granularity is presented on Table 1. It summarizes the three dimensions and scales of the monitoring space. Let us consider particularly the sampling granularity scale.

The sampling granularity depends on the tasks arrival process. A short burst of high utilization can cause saturation and performance issues, even though the overall utilization is low over a long interval. A major lead in balancing between the need of higher sampling frequency and reduction of the monitoring overload can be the Nyquist–Shannon sampling theorem: "If a function $x(t)$ contains no frequencies higher than B hertz, it is completely determined by giving its ordinates at a series of points spaced $1/(2B)$ seconds apart" [10].

Monitoring space			
❶ Monitored parameters list (list cardinality χ)		❷ Sampling granularity σ	❸ States precision π
Low level (resources)	High level (services)	Equals the ratio between Sampling periods and a Performance or Load parameter – Nyquist-Shannon sampling theorem	Equals the number of distinguishable states (or the number of the thresholds $\pi = \tau + 1$)
• Utilization, • Idle time, • Capacity • Transfer rates • Success and failure rates • etc.	• Rates • Latency and Queue time • Concurrency or Queue length • Success and failure rates • etc.	Fine sampling $\sigma \geq 2$ . . Semi-fine (medium) sampling $\sigma = 1$ . . Coarse sampling $\sigma < 1$	2 3 . . <scale> . . 100 (full 0-100% scale)

***Table 1**. The three-dimensional monitoring space.*

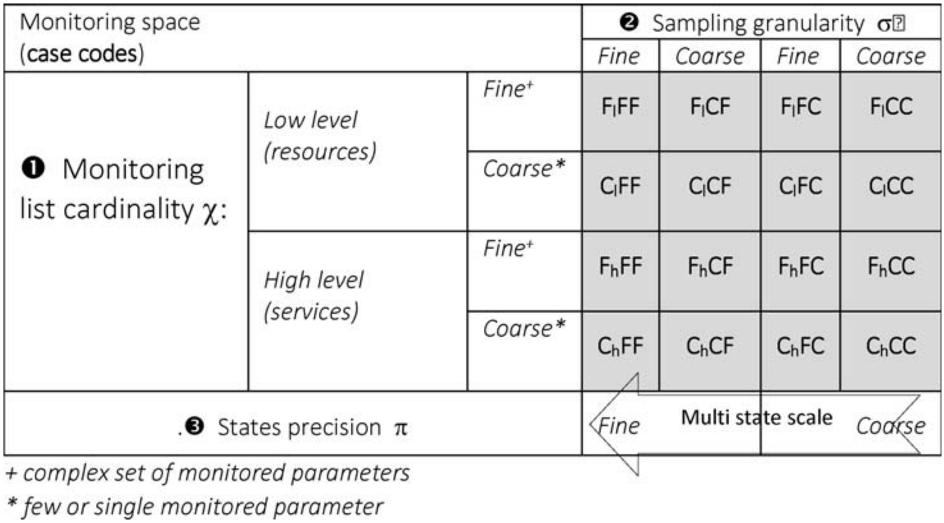
By the monitoring process, we have to consider the service process state as a function and consequently we can adapt the monitoring frequency to one of the following phenomena:

a) resources – the mean period between the transition of the resource state from one state to another – e.g. from normal to overloaded state. Here we see a typical example of the implicit correlation between sampling granularity σ and states precision π. If the system monitoring recognizes numerous resource states then the probability of interstate transition of the resource and consequently the frequency of the transitions is bigger. Following the Nyquist requirement one has to rise the sampling rate making σ finer.

b) services load (tasks) – the mean service time of the tasks and the mean time between the tasks' arrivals.

Both those levels of monitoring granularity require adaptation of the monitoring frequency to the current state of the service process.

For reference purposes it is continent to introduce codes of the monitoring granularity cases. We define these codes in Table 2. The codes do not provide precise value of the monitoring granularity triple <cardinality, sampling, precision>. Instead the quantitative values in <χ, σ, π> are replaced with qualitative symbols **C** for coarse and **F** for fine. Just an addition for the first code is the lower index *l* or *h* for the monitoring level – low or high respectively.



Monitoring space (case codes)			❷ Sampling granularity σ			
			Fine	*Coarse*	*Fine*	*Coarse*
❶ Monitoring list cardinality χ:	*Low level (resources)*	*Fine*$^{+}$	F_lFF	F_lCF	F_lFC	F_lCC
		*Coarse**	C_lFF	C_lCF	C_lFC	C_lCC
	High level (services)	*Fine*$^{+}$	F_hFF	F_hCF	F_hFC	F_hCC
		*Coarse**	C_hFF	C_hCF	C_hFC	C_hCC
.❸ States precision π			*Fine*	Multi state scale		*Coarse*

+ complex set of monitored parameters

** few or single monitored parameter*

Table 2*. Case codes of the monitoring space*

Let us consider how this coding works.

Taking the state precision on **$*_l*F \leftrightarrow *_l*C$** scale one has fine-grain to coarse-grain of the resource (for lower index l) monitoring. For **$*_l*F$** this means to divide the state space of a resource component to a relatively big number of distinguishable states. The ultimate practical granularity is 0-100% scale with its 100 distinguishable states. Obviously it is not very practical to monitor the actual utilization of a resource with such a fine precision. A more balanced still fine-grain monitoring frequency can distinguish several levels of underloading and several levels of overloading. E.g. a 6-thresholds scale can comprises the 7 states with following exemplary limits: idle (0%), near idle (1-5%), underloaded (6-20%), normal (21-70%), overloaded (71-90%), near busy (91-95%), bottleneck (96-100%). The information about the limit states – idle and bottleneck is – considered more valuable. E.g. in case of multiple resource components a newly arriving task can be planned on an idle or near-idle resource rather than on an underloaded resource. The same stands for the upper part of this scale[2]. These are $*_{\mathbf{l}}$***F** -scales. In comparison a $*_{\mathbf{l}}$***C**-scale would comprises just few thresholds – e.g. underloaded-overloaded states (1 threshold) or underloaded-normal-overloaded states (2 thresholds).

By analogy a $*_{\mathbf{h}}$***F** ↔$*_{\mathbf{h}}$***C** scale has fine-grain to coarse-grain service monitoring. This means to divide the state space of a service again to a number of distinguishable states but using high-level monitoring parameters. The QoS is usually described by latency time T_L and service time T_S. T_L and T_S are connected by the waiting or queue time T_S in the dependency:

$$T_L = T_Q + T_S \tag{1}$$

In order to use T_L as a local load measure one has to either use:

a) a benchmark service with known T_S and to keep record log of its consecutive T_{Li} – this is an easier and straightforward approach but generating additional overload

b) any service – by checking the ratio ρ between its service time T_S and queue time T_Q, or the queue time alone:

$$\rho = T_Q/T_S. \tag{2}$$

Again depending on monitoring granularity we may have various number of distinguishable states of the monitored service performance. In such an event scale for

$T_Q = 0$ (also $\rho = 0$ regardless T_S) one has idle resource. (3)

The upper part of the scale is context dependent on the load granularity itself

[2] The upper part of the utilization scale is of major use if the planning and enacting processes are designed to transfer partially executed tasks form busy and bottleneck resources or services to the less loaded ones. Such planning algorithms have far less application chiefly because of the overload of transferring partially executed tasks together with their context; e.g. a transfer of a virtual machine to another server would require to stop and store all local processes and their messages at proper checkpoints along with VM core itself, further to transfer the stored state to another node and restore the processes in reverse order. Planning the newly arrived tasks do not incur such overload. As a linear dependency reaches 100% bound of U it is suitable for real time servicing while the sigmoid line do not need to reach 100% for higher. Of course this curves are to be considered merely as numerical approximations or hypotheses and model to the real process of service monitoring – Fig. 2.

i.e. on T_S. For this upper part the ratio between the service time and queuing time can be used as measure of the system readiness. Analogously to an early concept for the *perceived time* in [1] we can estimate

$\rho = 1$ as indicator for 50% utilization (4)

of the servicing resource as half of the service latency is being spend in waiting line – Fig. 2. Not so obvious is the definition of which value of r represents 100% or near100% utilization of the servicing resource. If we want to keep *a linear scale* it is easy to define

$\rho \geq 2$ as indicator for 100% utilization. (5)

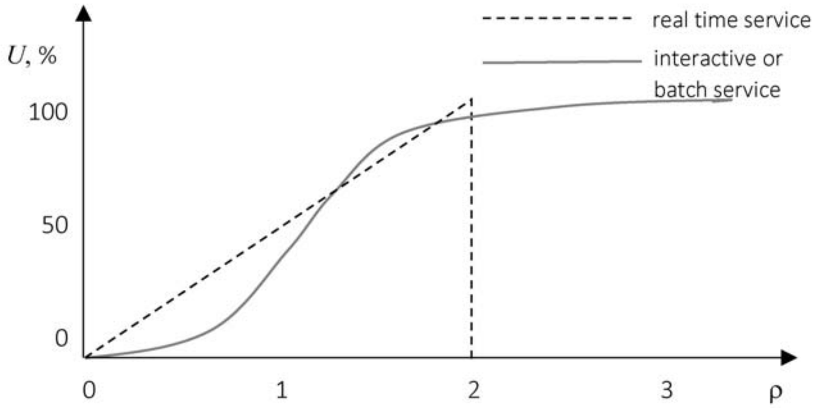


***Figure 2**. Utilization U as a function of service ratio* ρ.

One is not limited to this upper bound and to linearity as well. As a speculation on this subject we can suggest trigonometric sigmoid instead of a linear dependency $U(\rho)$ – Fig. 2. For example in the real time servicing the latency T_L is limited by one- or two-parameters deadline. In the interactive services T_L is bound to the specific QoS requirements. In the batch processing systems there are no practical limits to T_L. Consequently expressions (4, 5) are only exemplary as they are model- or assumption-dependent.

By $*_{\mathbf{h}}*\mathbf{F} \leftrightarrow *_{\mathbf{h}}*\mathbf{C}$ scale we may define coarser or finer number of states as for $*_{\mathbf{l}}*\mathbf{F} \leftrightarrow *_{\mathbf{l}}*\mathbf{C}$ monitoring scale. Let us consider these two scales by the examples of their typical monitoring set of parameters in the following section.

3. Case Study: Metrics Parameters

To illustrate how our monitoring granularity scheme reflects the “real-world” system metrics we will map several typical metrics sets taken from [3, 4, 8, and 9] on it.

Metrics capture a state of a property of a system or systems at a specific point in time — for example, the number of users currently logged in. They are usually collected at regular intervals to monitor a system over time. Let us consider the possible metrics:

3.1. Service (or Load) metrics

Load metrics show the high-level state and health of a system by measuring its visible or useful result. Although a load process can be described independently to the target system (e.g. by the number of tasks' operations and the tasks' interarrival time), it is more practical and applicable to measure a load process by the times of its servicing by the target server system. Load metrics are generally classified into several sub-types.

1. Throughput is the amount of usable work the system is doing for certain time period. Throughput is the rate at which a system completes operations, in units of operations per second.

2. Success metric represents the percentage of work that was executed successfully.

3. Error metric capture the number of erroneous results, usually expressed as a rate of errors per unit time or normalized by the throughput to yield errors per unit of work. Error metrics are captured separately from success metrics because errors could indicate serious problem with.

4. Concurrency is the number of operations in progress at a time, either as an instantaneous measure or an average over an interval of time.

5. Latency T_L is the most common performance metric, which represents the time required to complete a unit of work. It is the total time operations require to complete, from the perspective of the client (be it a user or another system). Latency is the time between making the request and getting the response. Latency is usually composed of two parts:

Queue time T_Q is the first component of latency: the time the request spends waiting, queued for service, after the request is made but before the work begins.

Service time

$$T_S = T_L - T_Q \tag{6}$$

is the second component of latency, after the device accepts a request from the queue and does the actual work. While T_L and T_Q depend on the local resource condition, T_S is independent to any resource parameter; it is just one of the two major load model parameters (the other is the interarrival process e.g. mean interarrival time).

Subtype	Description
throughput	http requests per second
success rate	percentage of http requests since last measurement
error rate	percentage of http requests returning http error code 500 (Internal Server Error) and/or unauthorized error since last measurement
latency95 (i.e. by 95% utilization)	95th percentile query time [mS]
latency99	99th percentile query time [mS]

***Table 3**. Example load metrics: Http web service – case code* **$\mathbf{F_h}$FF**

3.2. Resource metrics or performance

Resource metrics includes all physical server functional components (CPUs, disks, communication busses). Some resources are low-level — for instance, a server's resources include such physical components as CPU free or busy capacity, memory, disks, and network interfaces.

1. Utilization *U*: most of the time means the average percentage of time that a resource is busy performing useful work and for the various type of memories, utilization is the capacity of the resource that is used.

Even if the resource is fully busy, it can still accept more work - which is often put in a queue or makes the length of the system queue to grow instead of to shrink or stay stable. The degree to which it cannot do so is identified by saturation. Once a capacity resource reaches 100-percent utilization, no more work can be accepted, and it either queues the work (saturation) or returns errors. 100% utilization usually indicates bottleneck.

By coarse granularity the utilization is measured over a relatively long time period (multiple seconds or minutes), a total utilization of, say, 70% can hide short bursts of 100% utilization.

Some system resources, such as hard disks, cannot be interrupted during an operation, even for higher-priority work as opposed to CPUs, which can be interrupted ("preempted") at almost any moment. Once utilization is over 70%, queueing delays can become more frequent and noticeable.

In general for cost-effective cloud management targets no more than two situation – preventing of underloaded states of the system components as well as overloaded states. There is no much sense or QoS gain in load balancing between resources that are working with 40-50-60% utilization. Furthermore load balancing and leveraging schemes are applicable only if there are replicated or multiplied resources or services – depending on the control level.

Definition of utilization can be space- or time-oriented or both. This depends on the resource type. By I/O resources (e.g., disks) – utilization is the busy time. By capacity resources (e.g., main memory) – utilization is space or capacity consumed. Storage devices can be measured as both space- and time-consuming resources. Utilization varies during the service time. It is practical to measure it as cumulative percentage over a time interval (e.g., CPU is running at 90-percent utilization). This time interval or the period between the consecutive utilization values represents the granularity of the utilization monitoring. Generally the shorter sampling period means finer granularity. However, the concept of monitoring granularity is context dependent on the working mode of the resource. If the utilization of the resource stays stable during a long period of time then even a low frequency sampling can be interpreted as fine granularity. On the contrary, if the utilization fluctuates heavily then we need very short sampling period in order to monitor and eventually to control the service process.

4. Conclusion

The proposed monitoring granularity space defines its complexity as a three-component parameter which we call granularity. One can refer to the each scale of this monitoring space in order to compare various monitoring polices. The components of granularity have comparable and typically quantitative values. Each of the three scales – on χ, σ and π – is multi-state scale. Nevertheless we split these scales into two parts – “fine” and “coarse” which produces the 3-letter codes for the monitoring granularity

It is always practical to reduce a multicomponent estimation to single parameter. This is possible if we consider each of the parameters not merely as fine or coarse but as a full-scale parameter. Formally it is easy to propose a weighted scalar γ out of three parameters granularity $\Gamma(\chi, \sigma, \pi)$ provided we assume the same weight in the estimative formula:

$$\gamma = \sqrt{\chi^2 + \sigma^2 + \pi^2} \qquad (7)$$

However such straightforward approach does not take into account potentially different significance of the three components for the complexness of the monitoring granularity. A proper value of γ should serve as estimative of the system overload which a monitoring policy induces in the servicing infrastructure. Therefore the validity of (7) must be checked and possibly revised after extensive testing or benchmarking of the monitoring overload in whole scale presented in Table 1.

Acknowlegemets. This work is supported by the project ДН 02/9/17.12.2016. of the Bulgarian Science fund.

Refferences

[1] Geist, R. M., and Trivedi, K. S., "The Integration of User Perception in the Heterogeneous M/M/2 Queue," Proc. of the 9th IFIP Int. Symposium on Computer Performance Modeling, Measurement, and Evaluation (PERFORMANCE '83), College Park, Maryland, May, 1983, pp. 203-216.

[2] Georgiev, V. Numerical Solution of Cloud Servicing Models. Proceedings of the 2014 International Conference on Mathematics and Computers in Sciences and Industry (MCSI 2014) Varna, Bulgaria, 13-15 September 2014, pp 22.-26. ISBN 978-1-4799-4324-1/14

[3] Gregg B., "Thinking Methodically about Performance" Communications of the ACM, Vol. 56, No. 2, February, 2013, pp. 45-51.

[4] Jayathilaka H., Chandra Krintz, Rich Wolski, Performance Monitoring and Root Cause Analysis for Cloud-hosted Web Applications. WWW 2017, April 3–7, 2017, Perth, Australia. ACM 978-1-4503-4913-0/17/04.

[5] Kuchen, H., A. Wagener. Comparison of Dynamic Load Balancing Strategies. Parallel and Distributed processing. K. Boyanov ed. Elsevier Science Publishers. 1991. pp. 303-314.

[6] *Kwiatkowski, Jan (9 September 2001)*. "Evaluation of Parallel Programs by Measurement of Its Granularity". *Parallel Processing and Applied Mathematics. Springer Berlin Heidelberg: 145–153.* doi*:*10.1007/3-540-48086-2_16]

[7] Lamport, L., Time, clocks, and the ordering of events in a distributed system. Communications of the ACM. XXI, 7., 1978, pp. 558 - 565.]

[8] Li J., Naveen Kr. Sharma, Dan R. K. Ports, and Steven D. Gribble. Tales of the Tail: Hardware, OS, and Application-level Sources of Tail Latency Proceeding, SOCC '14 Proceedings of the ACM Symposium on Cloud Computing, Seattle, WA, USA — November 03 - 05, 2014.

[9] Nishtala R., H. Fugal, S. Grimm, M. Kwiatkowski, H. Lee, H. C. Li, R. McElroy, M. Paleczny, D. Peek, P. Saab, D. Stafford, T. Tung, and V. Venkataramani. Scaling memcache at Facebook. In Proceedings of the 10th USENIX Symposium on Networked Systems Design and Implementation (NSDI '13), Lombard, IL, USA, Apr. 2013.

[10] Nyquist–Shannon sampling theorem, https://en.wikipedia.org/wiki/Nyquist%E2%80%93 Shannon_sampling_theorem]

[11] Performance Visibility for Cloud. White Paper, SevOne https://www.sevone.com/white-paper/monitoring-cloud-infrastructure-performance-eliminate-visibility-gaps].

Optimizing and Verifying Model Effectiveness Using Random Forest and Logistic Regression

Vladimir Nasteski[1], Violeta Manevska[1], Snezana Savoska[1],

[1]Faculty of Information and Communication Technologies
University „St.Kliment Ohridski“ – Bitola,7000,ul. Partizanska bb, Macedonia,
v_nastey@hotmail.com, violeta.manevska@gmail.com, snezana.savoska@fikt.edu.mk,

Abstract. In the last decade, the volume, variety, velocity and the veracity of the information including the variety of devices that transmit the information rapidly increased and made a huge data flooding, creating big unstructured databases. Big Data refers to databases that vary by type, volume, velocity and variety. The standard techniques for data analyzing and optimizing fail to deal with these databases and it is necessary to find other methods and tools to deal with these problems. One of these tools is Apache Spark which has become one of the most popular tools when analytics and visualization of Big Data is taken into consideration. In this paper, the MLlib library is used as a part of Spark, for creating and analyzing data models. The two models are created using the Random Forest regression and Logistic Regression algorithms. Using these algorithms, the strong features that define the targets are presented and the models' effectiveness is verified.

Keywords: Big data, Spark, MLlib, Random Forest regression, Logistic regression.

1. Introduction

The recent interdisciplinary area that uses scientific methods, processes and systems for extracting knowledge and making deep data insight is Data Science. Today's data flood has various forms, unstructured or structured [1]. The data scientists combine statistics and mathematics, programming methods and newest problem-solving algorithms to deal with data complexity and to find unknown patterns in data. This is a great challenge for data scientists. The Big data concept is used to delve with heterogeneous data that cannot be analyzed or processed with traditional tools and methods. The gained Data Science knowledge for data variety, volume, velocity and veracity can be used by data scientists to develop a new technique that can be used for Big data processing, data visualization and data analytics.

The Data Science knowledge is used by many organizations in order to gain additional advantage for their organizations and to influence on some business improvement, taking into consideration that their future depends on the information and knowledge gained through data analysis [3]. Focusing on using some services that follows users' experience, compels business organization to use emerging software tools that use machine learning algorithms [4].

The increased Big Data analysis popularity is the main reason for usage of these novel tools and techniques. The machine learning algorithms is one of the techniques that provides capabilities to solve high complex problems and it is a desirable tool for data scientists [5] [31]. Such analysis leads to one of the most powerful uses of Big Data such as creation of prediction for the given set of responses.

A large set of machine algorithms is used for classification and regression. The Random Forest algorithm also belongs to these groups, although it is described mostly as a combination of predictors. The predictor's algorithm is influenced by a random vector value and independent samples through the same data distribution. For each predictor, the algorithm creates the tree in the forest [7]. The structure of the algorithm leads to low error rate. Also, one of the main uses of Random Forest algorithm is testing through many simulations, created for real problems [10]. This method can be used in applications with correlate predictors, even if the number of predictors is greater than the number of observations [9].

The Logistic regression algorithm is an approach that is used when the categorical depended variable can have only two values, as true or false [29]. As a method used for analyzing categorical responsive variables, for many data scientists this type of regression is better suited for modeling, compared to the discriminant analysis [13].

The variety of tools available for data analysis empowers developers to deal with data, to dig deeper into the data [14]. One of the most popular framework is Apache Spark [2], fast and cluster computing system that provides high level API's in Java, Scala, Python and R. As an optimized engine, Spark supports general execution graphs. The framework also provides faster and more general data processing platform which enables increasing of speed 100x, when the program run in memory, or 10x faster when run on disk than what Hadoop does [3]. Data scientists also use the framework for rapid scalable data transformation and data analysis. The framework supports set of high level tools as Spark SQL, MLlib for machine learning, GraphX for graph processing and Spark Streaming.

The paper aims to create, analyze and optimize models using both, Random forest regression and Logistic regression algorithms on the same dataset in the framework that provides scalability with a minimal fault tolerance. Using a visual Spark application, an open source dataset is implemented and then, the data are transformed to get reliable results. The model is evaluated using regression evaluation metric. In the end, the results of the model effectiveness are presented for the first model, and the results of the feature importance from the second model.

The paper is organized as follows. Section two highlight the relevant related work in the area. Section three provides an overview of mentioned algorithms, as well as the testing environment that is used. Also this section briefly describes the main steps for creating the models. Section four discusses the process of models' training using the algorithm Random Forest and Logistic Regression. In section five, the results of the model analysis are presented. Finally, some considerations and conclusions as well as future work are presented.

2. Related works

The general knowledge for Random Forest algorithm in the machine learning area is taken into consideration by Brieman [24]. The author combined different trees decision methods. Many authors explain improvement in the classification and regression accuracy, achieved by usage of the trees [8],[15]. In this model, each tree is built according to some random parameter. Later, the algorithm is widely accepted with successful application for general use in the classification and regression method.

This research paper motivates many scientists to dig deeper into algorithm, as is shown in Dietterich paper [18]. The algorithm is improved by using random subspace method with randomizing the internal decisions of learning algorithm. More improvements of the Random Forest algorithm are made by random split selection methods [19].

The next years, this algorithm evolved from simple algorithm to model's framework [21]. The author analyzes the Brieman's algorithm in details and proves that the convergence rate depends only on the important features that are used in the algorithm.

Zachari and Fridolin present technical details of how the algorithm behaves in theory and practice, using examples of the American politics literature [11]. The authors introduce a software, used for algorithm implementation as well as methods that are discussed in the paper.

The general article provided by Peng, Lee and Ingersoll enable for researchers, editors, and readers to give a set of guidelines that enable some answers to the question what to expect when the article uses logistic regression techniques [22].

Deeper and wider analysis of the Logistic Regression is enabled by the authors proposing a performance measure that can be estimated, taking into consideration positive and unlabeled examples used for evaluating model performance. The measure that they propose, can be used with a validation set in order to select regularization parameters for logistic regression [16]. In the next paper the authors introduce a logistic model for data matching and describe the corresponding odds ratio formula [23].

3. Model development

The process of models development is described in this section. The models are also tested, one of them is evaluated using the Random Forest Regression. In the beginning, the basics of the Random forest algorithm processes are also described. Next, the testing environment is given, providing step-by-step explanation for the models development.

3.1. Random forest

The random forests algorithm is created to merge the prediction of a couple of trees, trained in isolations [15]. The trees are individually trained and the predictions of the trees are combined using the averaging process. The main choices that have to be done in the process of building random tree are: the method for splitting the leaves, the type of predictor that will be used in each leaf and the method for randomizing the trees [21]. Each of these methods demands appropriate selection of the split candidate and method for evaluation of quality for each candidate. The most common choices use axis with aligned splits or linear splits.

This process begins with generating candidate splits, whereby the criterion for selection between the splits is evaluated. One of the most common approaches is the process of choosing the right split candidate that is developed by the purity function over the leaves.

The most common choice for predictors in each leaf is using the average response over the training points that took part in the leaf. The process of inserting randomization can be achieved in different ways. One way is the randomization of choosing the dimensions that can be used as split candidates for each leaf. Another way is the choice between the coefficients for random combinations of features. Disregarding which randomization is used, the thresholds can either be chosen randomly, or by optimization of some part or whole data set included in the leaf.

3.2. Logistic regression

The logistic regression classifier is similar to the linear classifier that uses the calculated scores for predicting the target class. This regression is a statistical method used for analyzing of dataset where the data of relation between the categorical dependent variable and one or many independent variables by estimating probabilities using a logistic function are measured [12].

The output is measured using a dichotomous variable in order to predict a binary output given as a set of independent variables, meaning that the logistic regression is linear when the outcome variable is categorical and it predicts the probability of occurrence of event by data fitting to a logit function.

This algorithm generates the coefficients for some features of interest in order to predict the logit transformation of the probability of presence:

$$logit(p) = b_0 + b_1X_1 + \cdots + b_kX_k$$

where p denotes the features of interest presence probability. The logit transformation is defended as follow [28]:

$$odds = \frac{p}{1-p} = \frac{probability\ of\ presence\ of\ charateristric}{probability\ of\ absence\ of\ characteristic}$$

$$logit(p) = \ln\left(\frac{p}{1-p}\right)$$

Estimation made in this regression give the parameters maximizing the likelihood of observed simple values, rather than parameters minimizing the squared errors sum.

3.3. Testing environment

The application used for testing purposes presented in paper is Seahorse. This is a visual framework that provides creation of custom, simple and visual Apache Spark models [25]. The available methods in Apache Spark are presented as Seahorse objects. With selection of a set of objects and tools, the model is defined and then it can be used for analyzing and visualizing different data types through application's visual interface [6]. These objects also create workflow for model's visual presentation.

The main advantage is that the created models run through web browser, such as Chrome or Mozilla. The application can be linked or connected to some cluster, such as YARN [26], Mesos [27] or Spark Standalone.

3.4. Data description

Apache Spark is created to support and load various data sources as local data sources or servers' databases, cloud databases as HDFS data system or Google Drive [17]. This application can also be connected with different files types with a wide range of databases from a simple mySql tables, to large CSV or JSON files. If the database is attached to the server, the user has to be aware of the speed of database reading, database format etc. Before data reading, the database has to be transformed - to be cleansed, in order to provide accurate custom calculations.

For this purpose, as open source dataset, "Supplier Directory Data" is used [20]. The dataset is presented in .csv format and includes a list of suppliers that indicates the supplies carried at that location and the supplier's Medicare participation status in the United States. With its 174 MB size, it contains

9 columns (Country, DBA Name, Address, City, State, Zip, Phone, Product Category Name, Competitive bid) and 716.681 rows.

As it was mentioned before, some additional preprocessing activities have to be undertaken. The additional preparation of the database has to be made in order to improve the raw data quality. The ReadDataFrame method is used for data loading and data reading.

3.5. Data transformation

The next step of the model creation is the additional data processing. If the data type is string, the value has to be encoded into integer in order to get the index from the string column. For this reason, the StringIndexer method is used. This method transforms the labels with string values to labels with indexes. The indexes are ordered by frequency of the labels. The most frequent label gets an index 0, etc. For this model, all of the string columns are indexed (column range 0-8). Using this method, new indexed columns are created with the prefix *indexed_*.

In order to apply calculations to the String indexed columns, the OneHotEncoder method is used to create binary vector from the indexed data. The method translates ordinal values to vector having “1” only at position given by input numeric value. In this step, the column range is 9-16, excluding the *Competitive_Bid* column.

For predicted results comparition, weather the competitive bid is true or false, the indexes of the *Competitive_Bid* column are used to create a new column - *Competitive_Bid_label.* In this column, data of the competitive bid - true get index 1 and the competitive bid - false get index 0.

For additional data analysis, model optimization and model training, the method VectorAssembler is used. With this method the column range 9-16 from dataset join single vector column named *features*, as shown on Table 1.

Following the steps described above, the dataset is ready for additional optimization in order to apply the Random Forest algorithm for the first model and the Logistic regression algorithm for the second model.

Table 1: Data preparation for the models

Competitive_Bid	***Competitive_Bid _label***	***features***
false	0	(23306, [(227, 1), (3450, 1), (6718, 1), (12577, 1), (14195, 1), (15373, 1), (18157, 1), (23304, 1)])
true	1	(23306, [(175, 1), (3228, 1), (6724, 1), (12368, 1), (14196, 1), (14794, 1), (18195, 1), (23300, 1), (23305, 1)])

4. Model training

In order to perform an evaluation of the models, data have to be spitted into testing dataset and training dataset. This operation can be done using the Split method. In both models, the data is split with 0.75 ratio. In order to train the model, the Fit method is used. In the first model, the RandomForestRegression method is used to define the strong features defining competitive bid, sorted by priority and evaluated by the Root mean square error metric. In the second model, the LogisticRegression method is used to distinguish whether the competitive bid is true or false. The process development is shown on Figure 1.

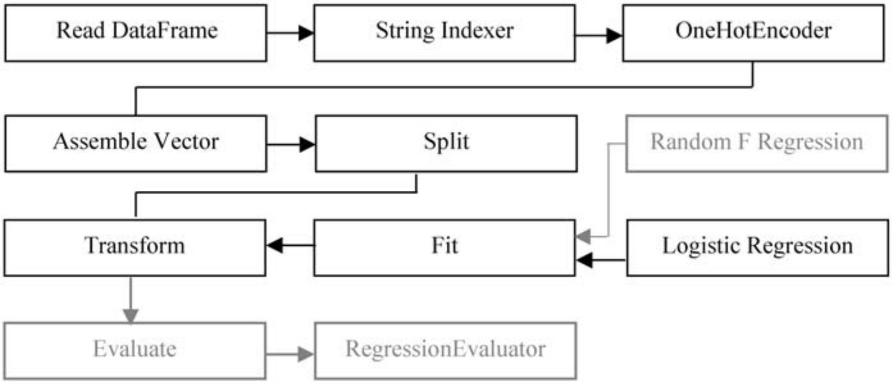


Fig. 1: The processes of Models' development

In the first model that uses Random forest regression algorithm, the method contains parameters that have to be modified to get the needed features' importance level. The parameters are: max depth, max bins and number of trees. There are many other parameters that can be tuned as: Maximum Depth, Maximum Bins, Minimum instances per node, Maximum Memory, Cache node IDs, Checkpoint interval, Regression impurity, Sampling Rate, Seed, Number of trees and Feature Subset strategy. For this model, they have default values. The input column in this method is *competitive_bid label*, and features column is vector column *features*. The max depth is 5, max bins 32 and the number of trees is 20. Also, in the end, the model can be evaluated using the *Root mean square error*, *R2* or *Mean absolute error.*

In the second model, the LogisticRegression method is used. The method has many parameters that can be tuned, for instance, elastic net mix parameter (for alpha = 0, the penalty is an L2 penalty and for alpha = 1, it is an L1 penalty) is set to 0, the regularization parameter (range constraints: 0 <= "regularization param" and "regularization param" is a floating point number), is set to 0, the tolerance parameter, etc. As an input column in the method, *competitive_bid_label* is used and vector column *features* as features column.

Further, both of the testing sets are ready for verifying effectiveness of the model. In order to get prediction using the training model, the Transform method is used. The effectiveness is evaluated by comparing the predicted data results from the model with the real data. These numbers are counted and presented as number.

5. Results and discussion

The model Fit returns a trained model. In the Random Forest regression model report, as an output the feature importance vector is calculated. The values from the n-th cell in this output indicates how much it affects the n-th feature on the overall prediction. The calculated values from the training model show the most important features defined in *competitive_bid_label*, sorted by descending order, shown in Table 2.

Table 2: Feature importance that define the Competitive Bid index

Feature	**Value**
State	0.0415439
City	0.000420416
Address	0.000284102

The results presented in Table 2 show that the feature "State" mostly affect the Competitive bid index in the Random Forest regression model. This value can be improved with increasing the numbers of tests and changing the parameters in the Fit object.

The Random Forest regression model is evaluated using standard deviation from the prediction errors using the Root Mean Square Error evaluation metric. The prediction error shows the maximal difference from the regression compared with the results. Taking into consideration calculated results from the model, the less significant aberration is shown with value of 0.121.

The output of the LogisticRegression method creates three new columns: *raw_prediction* (the confidence level), *probability* (column for predicted class conditional probabilities) and the *prediction* column created during model scoring. Using a simple SQL transformation, the values from the *prediction* column and the *competitive_bid_label* are compared. The model report shows no results where the *prediction* is 0 and the *competitive_bid_level* is 1, or where the *prediction* is 1 and the *competitive_bid_label* is 0, which means that the model performed well.

6. Conclusions

The aim of the paper is to highlight how the machine learning algorithms Random Forest and Logistic regression can be applied for deep analysis of certain problems through modelling, analyzing and visualizing datasets, using Spark. The main purpose of using machine learning algorithm is to create a model, test the feature importance, evaluate the results and verify the effectiveness of the model.

The general purpose of using the Random Forest algorithm is to calculate the strong features that describe the prediction feature, ordered by importance. In the Random Forest model, the strong features that describe the predicted feature Competitive Bid are: State, City and Address.

The calculated error rate generated from the evaluation metrics Root mean square is 0.121 that shows the maximum difference between the empirical value and the value calculated from the model.

After the comparison of the values gained from the prediction of the Logistic Regression model with the real values taken from the dataset, the conclusion is that the model has good performance, taking into consideration results which pointed zero falsely competitive bid values comparing to the real values from the dataset.

In the future, the model could be analyzed in details to minimize the error rate and detect the maximum deviation of some values. The higher deviations have to be optimized with model testing using additional values of the testing parameters. The models can be improved by fine tuning of the parameters of the algorithms that can lead to better optimization of the results.

References

1. Cárdenas, Alvaro A., Pratyusa K. Manadhata, and Sree Rajan. Big Data Analytics for Security Intelligence. *University of Texas at Dallas@ Cloud Security Alliance*, pp.1-22, 2013.
2. M. Zaharia, M. Chowdhury, M. J. Franklin, S. Shenker, and I. Stoica. Spark: cluster computing with working sets. *Proceedings of the 2nd USENIX conference on Hot topics in cloud computing*, pp.10-10, Boston, MA, June 22-25, 2010.
3. T. White. Hadoop: The definitive guide, 2nd ed. *"O'Reilly Media, Inc."*, 2010.
4. Alex Smola, S.V.N. Vishwanathan. Introduction to Machine Learning. *Cambridge University Press*, 2008.
5. Xiangrui Meng et al. MLlib: Machine Learning in Apache Spark. *Journal of Machine Learning Research 17, no. 1,* pp.1235-1241, 2016.
6. Karau, Holden, Andy Konwinski, Patrick Wendell, and Matei Zaharia. Learning spark: lightning-fast big data analysis - Ch11 Machine learning with MLlib. *"O'Reilly Media, Inc."*, 2015.
7. Gerard Biau. Analysis of a Random Forests Model. *Journal of Machine Learning Research 13*, pp.1063-1095, 2012.
8. L. Breiman. Consistency for a Simple Model of Random Forests. *Technical Report, UC Berkeley (670)*, 2004.

9. Andy Liaw, Matthew Wiener. Classification and Regression by random Forest. *R news*, *2*(3), pp.18-22, 2002.
10. Segal, M.R. Machine learning benchmarks and random forest regression. *Center for Bioinformatics & Molecular Biostatistics*, 2010.
11. Zachary Jones, Fridolin Linder. Exploratory Data Analysis using Random Forests. *Prepared for the 73rd annual MPSA conference*, 2015.
12. Fan, Rong-En, et al. LIBLINEAR. A library for large linear classification. *Journal of machine learning research,* pp.1871-1874, 2008.
13. Ng, A. Y., & Jordan, M. I. On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. *Advances in neural information processing systems,* pp. 841-848, 2002.
14. Tarwani, K. M., S. Saudagar, and H. D. Misalkar. "Machine learning in big data analytics: an overview." International Journal of Advanced Research in Computer Science and Software Engineering 5.4, pp.270-274, 2015.
15. L. Breiman. Random Forest. *Machine Larning 41(1),* pp.5-32, 2001.
16. Lee, Wee Sun, and Bing Liu. Learning with positive and unlabeled examples using weighted logistic regression. *ICML*. Vol. 3. 2003.
17. Google drive official website, https://www.google.com/drive/ (consulted on 18.03.2017)
18. Ho, T. K. The random subspace method for constructing decision forests. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp.832-844, 1998.
19. Dietterich, T. G. An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization. *Machine Learning,* pp.139-157, 2000.
20. A federal government website managed by the U.S. Department of Health & Human Services, https://www.healthdata.gov/ (consulted on 02.08.2017)
21. Criminisi, A., Shotton, J, and Konukoglu, E. Decision forests: A unified framework for classification, regression, density estimation, manifold learning and semi-supervised learning. *Foundations and Trends® in Computer Graphics and Vision*, *7(2–3),* pp.81-227, 2011.
22. Peng, Chao-Ying Joanne, Kuk Lida Lee, and Gary M. Ingersoll. An introduction to logistic regression analysis and reporting. *The journal of educational research,* 96.1, pp.3-14, 2002.
23. Steyerberg, Ewout W., and Marinus JC Eijkemans. Prognostic modeling with logistic regression analysis. *network* 10, p.11, 2000.
24. L. Breiman. Bagging predictors. *Machine Learning, 24(2),* pp.123–140, 1996.
25. Github repository, Scala version: 2.11.8+, Spark version: 2.0, Hadoop version: 2.7.0, https://github.com/deepsense-io/seahorse-workflow-executor (consulted on 21.02.2017)
26. YARN official website, https://yarnpkg.com/lang/en/docs/install/ (consulted on 16.05.2017)
27. Mesosphere official website, https://mesosphere.com/ (consulted on 12.05.2017)
28. Hosmer Jr, David W., Stanley Lemeshow, and Rodney X. Sturdivant. *Applied logistic regression*. Vol. 398. John Wiley & Sons, 2013.
29. Kleinbaum, David G., and Mitchel Klein. Analysis of matched data using logistic regression. *Logistic regression*. Springer New York, 2010. pp.389-428.

Author Index

Avdjieva Irena *54*

Bocevska Andrijana *80*

Dimeska Katerina *60*

Dimitrov Anatoliy *38*

Dimitrov Vladimir *91*

Dimitrova Monica *109*

Georgiev Vasil *136*

Hristov Hristo *136*

Kaloyanova Kalinka *7*

Kanabar Vijay *7*

Kleanthous Stella *17*

Kokkinaki Angelika *17*

Krachunov Milko *54*

Kulev Ognyan *54, 119*

Kyrillou Chrisostomi *17*

Manevska Violeta *149*

Milevski Dragan *69*

Milevski Ivan *80*

Nasteski Vladimir *149*

Patias Ioannis *31*

Peychev Deyan *45*

Savoska Snezana *60, 69, 80, 149*

Simeonov Daniel *136*

Simeonova Valeriya *125*

Vassilev Dimitar *38, 45, 54, 109, 125*

Zioga Fotini *17*