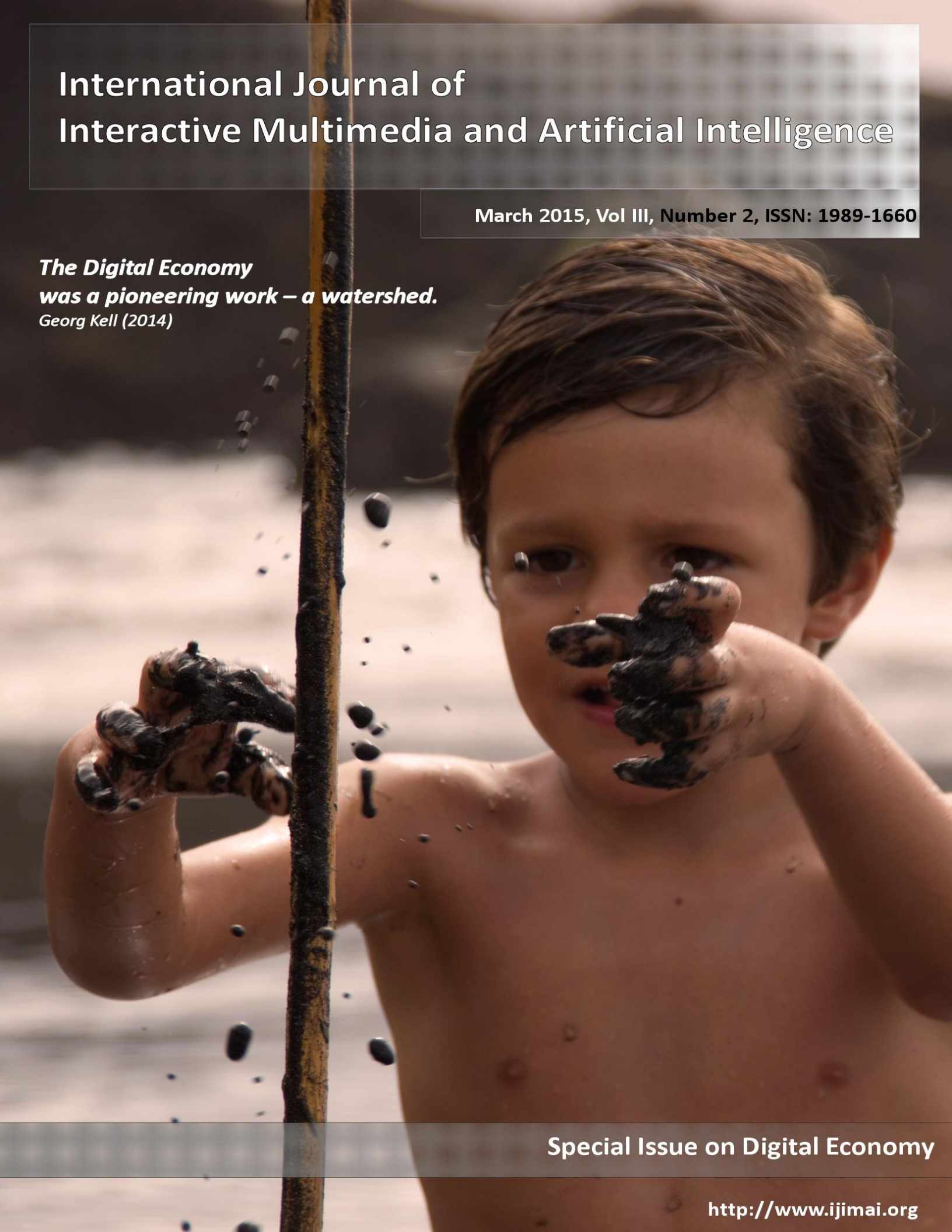
International Journal of Interactive Multimedia and Artificial Intelligence

March 2015, Vol III, Number 2, ISSN: 1989-1660

The Digital Economy was a pioneering work – a watershed.
Georg Kell (2014)



Special Issue on Digital Economy

http://www.ijimai.org

INTERNATIONAL JOURNAL OF ARTIFICIAL INTELLIGENCE AND INTERACTIVE MULTIMEDIA
ISSN: 1989-1660 - VOL. III, NUMBER 2

Editor's Note

The term 'Digital Economy' was coined for the first time by Don Tapscott in 1995 in his best-seller *The Digital Economy: Promise and Peril in the Age of Networked Intelligence*.[1] . When he wrote the book 20 years ago, he announced how he thought the Internet would fully transform the nature of business and government.

We have now extended the concept, illustrating how digital technologies are rapidly transforming business practices, the economy and societies. Technology, and its impact on business strategy and society, continues to rise in importance. The Digital Economy, sometimes also called "Digital Business" has become a philosophy for many top executive teams as they seek competitive advantages in a world of fast moving technological change. When we talk about digital technologies, we are not only talking about the internet, nor only ICT (Information and Communications Technology), but other concepts such as mobile, telecommunications or content.

The digital economy is by no means an exclusively economic concept. Therefore, it might be more appropriate to speak of digital society or digital technology. What matters is that digital is a transverse concept that affects individuals, businesses and public administrations.

People are progressively entering the digital world. In our daily life we are in direct contact with digital devices such as cars, electrical appliances etc. People increasingly communicate digitally; through mobile devices, internet, and social networks. In addition, our leisure, education and health are largely being integrated into a digital environment.

Likewise, organizations are progressively incorporating digital technology in their production and distribution processes, conducting research and also in the process of decision making. In this sense the emergence of Big Data has been one factor that has accelerated an already observed trend.

Public administrations and political institutions are incorporating digital into both their internal procedures and in their relationships with citizens. An example is the role played by Big Data in the two elections won by Obama. Another example may be the way the ISIS terrorist group uses social networking to publicize their activities and secure funding.

But digital technology and information are not only affecting the daily behavior of individuals, businesses and public administrations. They are also playing a very prominent role in the knowledge of the fundamental laws of nature. In this sense we can highlight the contribution of digital technology in the computer modeling of fundamental elements of life, and consequently in the creation of artificial life [2]. Likewise, you may also note the role in deepening the knowledge of the universe. [3]

Bearing in mind that Digital Economy is becoming more relevant from the economic perspective, we decided to star this issue with something unusual in this magazine: a paper that discusses the economy. It's interesting from our point of view, as it underscores the fact that digital technologies, which of course include Artificial Intelligence and Interactive Multimedia, are not only relevant from the technical point of view, but from the business and society perspective as well.

The paper *Differences in Measuring Market Risk in Four Subsectors of the Digital Economy* by Sonia Benito, Rebeca de Juan, Ricardo Gómez and Francisco Mochón, emphasizes two concepts: the Digital Economy is more than only ICT, (actually they identify four main sectors - Telecom Companies, Mobile/Internet Contents and Services, SW&IT Services and Application Software); and the Digital Economy is a heterogeneous sector, which is best to consider as the result of adding the four different subsectors.

The paper tests whether indeed the economic and financial performance of a portfolio of listed companies in each of the four subsectors presents relevant differences. In order to measure the risk, they use the average value at risk to estimate market risk of the four subsectors of the digital economy. The nature of the business models of these subsectors is different, each one presenting very different risk profiles. The riskiest subsectors are Mobile/Internet Contents & Services followed by SW&IT Services and Application Software. On the contrary, the Telecom sector is by far the safest. These results support the hypothesis that the Digital Economy is not a homogeneous sector.

It is estimated that currently two billion people are connected to the internet. The internet has transformed the way we live, the way we work, the way we socialize and meet, as well as the way our countries develop and grow. In less than two decades, the internet has changed from a network for researchers, scientist and geeks, to a day-to-day reality for billions of people which in itself is a defined part of business opportunity and interaction where individuals have taken it upon themselves to create and refine the digital age. Even though Social Interaction was not one of the subsectors previously commented, it has become essential in everything related to digital where its importance is above question. People have already mastered the use of Social interaction to the point where they demand them. To this end Companies are improving their mastery of them, as they have found out that using them they can enhance operation, exploit new markets or customers loyalty.

In paper *The Digital Economy: Social Interaction Technologies – an Overview* by Teofilo Redondo exposes how Social Interaction Technologies (SIT) has had a transformational effect in many aspects of our lives. Since they touch so many parts of our lives, they have impacted on so many fields, in a clear social and technical convergence. Daily

DOI: 10.9781/ijimai.2015.320

activities of many businesses are being socialized, incorporating the central themes of social software (create, connect, contribute, and collaborate) into a multidisciplinary ecosystem of interactive and networked computing.

The paper reviews a number of social interaction tools and some special use cases and their greatest impact. The economic results of the so-called social media economy have yet to be produced, but only in terms of increased productivity, employees and customers satisfaction, that certainly have remarkable value.

Content is an essential part of the Digital Economy and a crucial player in one of the four subsectors discussed in the paper: Mobile/Internet Contents and Services. One of the challenging issues that needs to be dealt with when working with content as a digital business, is that the content needs to be adaptable for a wide variety of devices and platforms, acceptable to the consumer and the business. (smartphones, tablets, web browsers, eBooks …). The paper *New challenges on crossplatform digital contents* by Jesús Iglesias Feijoo and Guillermo Amat Gomariz, reveals a new challenge as there are more devices and platforms available (e.g. watches, glasses, cars) and many more to follow. The Internet of things will transform the technological world in which we develop an amalgamation of devices and interfaces. According to a study by Cisco[3], in 2015 there will be twenty billion connected devices, and this figure could be doubled by 2020.

Based on the work done under the Visio Project, funded by the Spanish Ministry of Industry, Energy and Tourism, the paper analyses the challenge over the next coming years for getting all devices to communicate between each other, regardless of technology and platform. Ultimately, a truly universal platform to avoid market fragmentation and provide access to information and services is proposed.

Another important challenge, when working with content, is how the end user discovers it. On one hand, customers don´t always know exactly what they want and specifically what they are looking for, and on the other hand, there is huge content growth, available in their various content types (video, Apps, Websites, etc.) on the Mobile/Internet. Our culture and economy is increasingly shifting away from the focus on a relatively small number of "hits" with a big number in sales. This in turn to the development of niche items with a small number of sales. Anderson predicted that we are "leaving the age of information and entering the age of recommendation" [4].

To face these challenges, recommendation engines (RE) are becoming highly popular. An RE offers new items (products or content) to users, based on their profile and historical data. The most popular algorithms used in RE are based on collaborative filtering [5]. This technique makes recommendations based on the past behavior of other users and the similarity between users and items. RE bases their results on data analysis. The greater the amount of pertinent data, the more accurate the recommendation will be. And the application of different Data Mining algorithms, in the case of the next paper, the application of Social Network Analysis (SNA).

In the paper *Empirical Comparison of Graph-based Recommendation Engines for an Apps Ecosystem* by Luis F. Chiroque, Héctor Cordobés, Antonio Fernández Anta, Rafael A. García Leiva, Philippe Morere, Lorenzo Ornella and Fernando Pérez, Agustín Santos, evaluates the performance of several REs based on the properties of the networks formed by users and items. The REs use a novel approach to graph theoretic concepts like edge weights and network flow. The evaluation has been conducted in a real environment (ecosystem) for recommending apps to smartphone users. The analysis of the results allows for the possibility that the effectiveness of an RE can be improved if the age of the data and a global view of the data is considered. It also shows that graph-based RE is effective, but more experiments are required for a more accurate characterization of their properties.

The Entertainment industry is one of the biggest source of content and videogames as well as one of the most important and profitable within it. Even though videogames are considered content themselves, the creation process requires the generation of several other pieces of content. At the moment, the creation of a videogame is often a large-scale endeavor and bears many similarities with, e.g., movie production. On the central tasks in the development of a videogame is content generation, namely the definition of maps, terrains, non-player characters (NPCs) and other graphical, musical and AI-related components of the game. Such generation is costly due to its complexity, the great amount of work required and the need of specialized manpower, hence the relevance of optimizing the process and controlling costs.

Paper *Procedural Content Generation for Real-Time Strategy Games* by Raúl Lara-Cabrera, Mariela Nogueira-Collazo, Carlos Cotta and Antonio J. Fernández-Leiva, exposes how procedural content generation (PCG) comes in handy as a means of reducing costs by using algorithmic techniques to automatically generate some game contents. PCG refers to the algorithmic creation of content for video games possible, such as maps, levels, terrains, graphic textures, music, rules, quests, narrative, and missions among other things [6]. PCG also provides advantages in terms of player experience, since the contents generated are typically not fixed but can vary in different playing sessions, and can even adapt to the player. For this purpose, the underlying algorithmic technique used for PCG must be flexible and adaptable. This is the case of computational intelligence in general and evolutionary algorithms in particular. In this work we shall provide an overview of the use of evolutionary intelligence for PCG, with special emphasis on its use within the context of real-time strategy games. We shall show how these techniques can address playability and aesthetics, as well as improving the AI game.

Since the Internet opened a new way to communicate in

many different forms, several sectors have adopted it. One of the most relevant has been the educational sector, which adopted such technology and developed the Web-based Educational Systems (WBES). Today this has become even more popular with the eruption of MOOCs (Massive Open Online Course). WBES platforms store and manage huge amounts of data. The stores of data are growing exponentially and contain hidden information that could be very useful to the users (both teachers and students).

The paper *Mining Web-based Educational Systems to Predict Student Learning Achievements* by José del Campo-Ávila, Ricardo Conejo, Francisco Triguero and Rafael Morales-Bueno studies how data mining can be used to induce student models from the data acquired by a specific Web-based tool for adaptive testing, called SIETTE [7]. Consequently top down induction decision tree algorithms are used to extract patterns, because these models (decision trees) are easily understood. In addition, the validation processes conducted have assured high quality models.

Digital technologies can be applied to different sectors and healthcare has been always a sector that demands creative, thoughtful uses of technology in different areas: devices, diagnosing, as well as management. In the management area, the development of health information systems has been guided by the need to manage the huge amounts of information that make the use of physical methods unfeasible [8]. However, these systems are not usually constrained set protocols. The result is that different hospitals working together or even different services within the same hospital cannot share information about their patients.

The paper *A Repository of Semantic Open EHR Archetypes* by Fernando Sánchez, Samuel Benavides, Fernando Moreno, Guillermo Garzón, Maria del Mar Roldan-Garcia, Ismael Navas-Delgado and Jose F. Aldana-Montes describes a repository of openEHR archetypes [9] that have been translated to OWL (Ontology Web Language). In the work, five different CKMs (Clinical Knowledge Managers) have been downloaded and the archetypes have been translated to OWL. This translation is based on an existing translator that has been improved to solve programming problems within certain structures, as part of the repository, a tool that has been developed to keep it up-to-date. So, any change in one of the CKMs (addition, elimination or even change of an archetype) will involve translating the changed archetypes once more. The repository is accessible through a Web interface (http://www.openehr.es/).

Another application for healthcare technology, in this case, a variety of disciplines, are shown in the paper *Auto-adaptative Robot-aided Therapy based in 3D Virtual Tasks controlled by a Supervised and Dynamic Neuro-Fuzzy System* by L. D. LLedo, A. Bertomeu, J. Díez, F. J. Badesa, R. Morales, J. M. Sabater and N. Garcia-Aracil. It presents an application formed by a classification method, based on the architecture of ART neural network (Adaptive Resonance Theory) [10] and the Fuzzy Set Theory, to classify physiological reactions in order to automatically and dynamically adapt a robot-assisted rehabilitation therapy to the patient needs, with the aid of 3D tasks in a virtual reality environment. Firstly, the mathematical and structural model of the neuro-fuzzy classification method is described, together with a signal and a data training acquisition. The virtual task is designed along with its physical behavior and procedures. Finally, the general architecture of the experimentation for the auto-adaptive therapy, is presented using the classification method with the virtual reality exercise.

The last paper in this issue *GPGPU Implementation of a Genetic Algorithm for Stereo Refinement* by Álvaro Arranz and Manuel Alvar, explores the advantages of using GPGPU implementation to speed up a genetic algorithm used for stereo refinement. The main contribution of this paper is analyzing which genetic operators take advantage of a parallel approach and the description of an efficient state- of-the-art implementation for each one. As a result, speeds of up to 80x can be achieved, which is close to real-time performance.

Dr. Francisco Mochón
Juan Carlos Gonzalvez

REFERENCES

[1] Tapscott, Don (1997). The digital economy: promise and peril in the age of networked intelligence. New York: McGraw-Hill. ISBN 0-07-063342-8.W.-K. Chen, *Linear Networks and Systems* (Book style). Belmont, CA: Wadsworth, 1993, pp. 123–135.

[2] Craig Venter. J, et. alia. (2010) Creation of a Bacterial Cell Controlled by a Chemically Synthesized Genome. Science 2.Vol. 329 no. 5987 pp. 52-56.

[3] Joshua Filmer . (2013)"An Ultracold Big Bang: A successful simulation of the evolution of the early universo". December 25. http://www.fromquarkstoquasars.com/an-ultracold-big-bang-a-successful-simulation-of-the-evolution-of-the-early-universe/

[4] Goldman Sachs. Publications by Sector/Industry https://360.gs.com/gs/portal/research/industrylist.

[5] The Internet of Things. How the Next Evolution of the Internet Is Changing Everything. Dave Evans, Cisco, Cisco Internet Business Solutions Group (IBSG), April 2011. http://www.cisco.com/web/about/ac79/docs/innov/IoT_IBSG_0411FINAL.pdf

[6] C. Anderson, "The Long Tail: Why the Future of Business Is Selling Less of More," Hyperion, 2006.

[7] Y. Koren, R. M. Bell, "Advances in collaborative filtering," in Recommender Systems Handbook, Springer, pp. 145–186, 2011.

[8] N. Shaker, J. Togelius, and M. J. Nelson, Procedural Content Generation in Games: A Textbook and an Overview of Current Research. Springer, 2014.

[9] R. Conejo, E. Guzmán, E. Millán, M. Trella, J. L. Pérez-De-La-Cruz and A. Ríos, "SIETTE: A Web-Based Tool for Adaptive Testing", International Journal of Artificial Intelligence in Education, vol. 14, pp. 29-61. 2004.

[10] The socio-economic impact of interoperable electronic health record (EHR) and ePrescribing systems in Europe and beyond. http://www.ehr-impact.eu/downloads/documents/EHRI_final_report_2009.pdf

[11] Catalina Martinez-Costa; Marcos Menárguez-Tortosa; Jesualdo Tomás Fernández-Breis. An approach for the semantic interoperability of ISO EN 13606 and OpenEHR archetypes. Journal of Biomedical Informatics. October 2010.

[12] G. Carpenter, S. Grossberg, N. Markuzon, N. Reynolds, D. Rosen, "Fuzzy ARTMAP: A neural network architecture for incremental supervised learning of analog multidimensional maps", IEEE Transactions on Neural Networks 3(4), pp. 698-713, 199

TABLE OF CONTENTS

OPEN ACCESS JOURNAL

ISSN: 1989-1660

COPYRIGHT NOTICE

Differences in Measuring Market Risk in Four Subsectors of the Digital Economy

S. Benito, R. de Juan, R. Gómez and F. Mochón

Universidad Nacional de Educación a Distancia, ZED, Spain

***Abstract* — This paper defends the wisdom of not considering the Digital Economy to be one homogeneous sector. Our hypothesis is that it is best to consider it the result of adding four different subsectors. We test whether indeed the economic and financial performance of a portfolio of listed companies in each of the four subsectors presents relevant differences. We use the value at risk measure to estimate market risk of the four subsectors of the digital economy. The riskiest subsector is *Mobile/Internet Contents & Services* followed by *SW&IT Services and Application Software.* On the contrary, the *Telecom* sector is by far the safest one. These results support the hypothesis that the Digital Economy is not a homogeneous sector.**

***Keywords* — digital economy, ICT, market risk, value at risk, volatility**

I. Introduction

DESPITE the importance of the Digital Economy, people often have two misconceptions when talking about it. The first mistake is to identify Digital Economy with ICT (Information and Communications Technology) sector. The second error is to consider the digital economy a homogeneous whole.

When the digital economy is identified with ICT, we are obviating that there is a big ecosystem of companies that provide digital content and services [1]. Moreover, the sector of the digital economy is not homogeneous. As we shall see in the next section, the Digital Economy covers various subsectors that have quite different characteristics.

In this regard, this article has two objectives: first, justify the different sectors that make up the digital economy from a technological and economic perspective; and second, find an objective criterion that allows us to verify the different nature of the subsectors mentioned.

To achieve these objectives, this paper has been structured as follows. In the next section, the various subsectors that make up the digital economy are analyzed. In the third section, a criterion is selected to evaluate the different behaviour of the various subsectors. The chosen criterion is market risk and volatility of financial returns. The fourth section presents the empirical analysis and the results are discussed. Finally, the fifth section presents the main conclusions.

II. The Four Subsectors of the Digital Economy

For most of the twentieth century, the telecommunications industry and the information technology industry were two distinct areas of activity. With the advent of the Internet, their paths crossed to the point of constituting a single industry known as ICT (Information and Communications Technology). In the context of ICT, all types of information (voice, data, and images) could already be processed, stored and transmitted between devices, at any time and from any location.

Additionally, in recent years, the explosion of mobile Internet, along with the phenomenon of smartphones, tablets and social networks, has promoted the creation of a new economic sector: the services and digital content sector. Thus, there are new companies able to offer thousands of new products and services to billions of users through all kinds of networked devices. It has created a true 'digital revolution' whose impact is much greater than any previous technological progress because of its transverse and disruptive nature.

This revolution not only changes the way we work and communicate, as was the case through the traditional ICT industry, but also creates new business models. These business models break the foundations on which the economic activities that have historically accounted for most of the world's GDP are based.

This whole phenomenon of convergence between traditional ICT companies and companies which provide services and digital content through all kinds of networks set an ecosystem that is commonly called the digital economy.

Companies participating in this digital economy are diverse and heterogeneous, occupying a very different place within the industry value chain. In practice, this means that we should speak necessarily of various subsectors within the Digital Economy. Thus, taking into account the position of firms in the value chain, four subsectors of homogeneous firms can be identified [2] (see Figure 1).

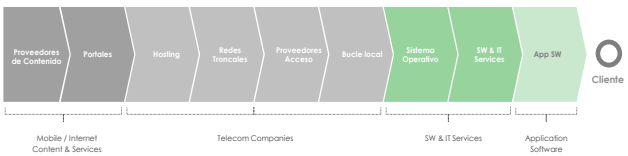


Fig. 1. The four subsectors of the digital economy

DOI: 10.9781/ijimai.2015.321

1. Telecommunications

Companies in this subsector are responsible for ensuring connectivity and allowing content and online services to be delivered from content providers to physical devices with which customers can enjoy such contents.

These companies can perform many activities within the value chain: hosting, management backbones, Internet access supply and access to the local loop / client device, etc.

Increasingly, the role of telecommunications companies is becoming a 'commodity' and it is difficult to differentiate their products and services from those offered by their competitors.

In addition to the risk of 'commoditization,' we find other risks such as financial risks linked to high investments to be made every few years in network equipment, transmission and security; the appearance of new competitors and suppliers like cable TV and satellite operators; or OTT[1] companies offering alternative services (e.g., WhatsApp, Skype, Waki, TV, Line, etc.) covered under the controversial principle of 'net neutrality.' There is also regulatory risk.

2. Mobile/Internet Content and Services

This subsector consists of companies that create or acquire content in various formats --text, audio, images, video, music, blogs, etc.-- and provide digital services that deliver value to their customers through the Internet (broadband networks and/or mobile), and can be enjoyed through many devices (PCs, smartphones, tablets, etc.).

In this subsector, two types of firms are grouped: i) companies that create and publish their own content, and ii) companies that add third-party content and publish it through a single Web/Wap portal. Also, these companies can distribute many types of content and services: news, entertainment, e-commerce, search engines, travel agencies, education, etc.

The main risks inherent to these companies deal with their innovativeness, ability to attract talent, rapid adaptation to changes in the habits and preferences of their users and the ability to monetize their products and services.

3. Software and Information Technology (SW & IT) Services

Companies in this subsector develop SW and technology that allow users to enjoy online content and services offered through Internet-connected devices. The objective of these companies is to isolate the user from the technological complexity and provide a good user experience.

These companies are often subject to the so-called 'network effect' whereby users of a specific type of service benefit from the increase in the number of users who use it, which often leads to a few companies having a very high market share.

Other features of the market in which these companies operate and significantly affect their risk are product decisions concerning standardization against the 'customization' of the product, the high costs of change / replacement of the product, and being subject to the law of 'increasing returns', which means that the costs of creating the proprietary software are very high while duplication or mass production costs are much smaller and are subject to strong economies of scale. This means that the marginal cost of producing an extra unit tends to zero as output increases.

4. Application SW

Such companies develop SW running on the Operating System (OS) of the device and help users to perform certain tasks, increasing their productivity.

This SW should be independent of specific hardware used and must be implemented without major problems in other devices and/or operating systems. It can also be distributed by building a 'bundle' with the operating system or be distributed independently.

The market in which the companies in this subsector operate also shares two characteristics of the previous subsector: the so-called 'network effect' and the law of 'increasing returns.'

Once we have submitted the four subsectors, we can formulate the following question: Is there any objective criterion that may ratify the existence of the four different subsectors?

From all the above, one might state that a different position of the companies from every subsector in the value chain of the digital economy would also imply a different degree of value added to the process and therefore a different risk for firms in each subsector.

Additionally, if the risk faced by enterprises is different within a subsector level, the target shareholder's return in each of the subsectors should also be. This fact (if true) should also be shown in the risk-return ratio inherent in the shares of listed companies in the digital economy.

To support this differentiation by subsectors, we will test whether indeed the economic and financial performance of a portfolio of listed companies in each of the four subsectors shows relevant differences.

More specifically, we will measure the financial risk of the four groups of companies representing the aforementioned subsectors. For this task, we will use two kinds of measures: (i) volatility, which is the traditional measure of risk, and (ii) Value at Risk, which is currently the most used.

III. MEASURING MARKET RISK: VOLATILITY - VALUE AT RISK

A context of risk is one in which we do not know with certainty the consequences associated with a decision. The only thing that we know is possible outcomes associated with it and the likelihood of achieving such results. In the financial field, the notion of risk implies that we know the various yields can potentially get to make an investment??? and also know the probability of achieving such results. This allows us to estimate the average expected yield and the possible diversion 'above' or 'below' the average value, that is, the risk. The most popular and traditional risk measure is volatility (variance). In fact, traditional financial theory defines risk as the dispersion of returns due to movements in financial variables.

Another way of measuring risk, which is the most commonly used at present, is to evaluate the losses that may occur when

[1] OTT, Over-The-Top, describes a scenario in which a telecommunications service provider delivers one or more of its services across all IPS.

the price of the asset that makes up the portfolio goes down. This is what Value at Risk (VaR) does. The Value at Risk of a portfolio indicates the maximum amount that an investor may lose over a given time horizon and with a given probability. In this case, the concept of risk is associated with the danger of losses. Since the Basel Committee on Bank Supervision at the Bank for International Settlements requires a financial institution to meet capital requirements on the basis of VaR estimates, this measurement has become a basic market risk management tool for financial institutions.

Formally, the Value at Risk of a portfolio at (1-α)% confidence level is the α quantile of the probability distribution of the return portfolio. For instance, suppose that the Value at Risk of a bank´s portfolio on the horizon of 1 day is -3.5% with a confidence level of 95%. This means that with a probability of 95%, the return of the portfolio will be higher than -3.5%. There is still a 5% chance that the return of the portfolio will be below -3.5%.

Although the concept of Value at Risk is very simple to calculate, it brings some difficulties.[2] Under the framework of the parametric techniques, the estimation of the Value at Risk requires a forecast of the portfolio return´s conditional volatility.

The volatility of the financial returns can be estimated using different models. In this paper, we use the beta-skewness-t-EGARCH model proposed recently by Harvey and Sucarrant (2013). This model captures some of the characteristics of the financial returns like (i) '*cluster in volatility*' and (ii) the '*leverage effect*'. The former means that large returns in absolute value are likely followed by other large returns in absolute value, and small returns in absolute value are followed by small returns in absolute value. The leverage effect means that volatility tends to be higher after negative returns; this is typically attributed to leverage (hence the name). For the estimation of the volatility model, we assume that the probability of the return portfolio going below zero (daily average return) is higher than the probability of going above zero, which is in keeping with the empirical evidence (skewness distribution).

To evaluate the accuracy of these estimations, we´ will use several standard tests (see Appendix A). In addition, we use Lopez´s loss function to measure the magnitude of the no cover losses. The losses are not covered when the portfolio return goes below VaR. For instance, at time *t*, the value at risk of a portfolio 1 day ahead is -3.5%. A day later, we observe that the return portfolio was -5.0%. In this case, 1.5% of the losses were not covered.

In the following sections, we use the value at risk measure to estimate market risk of the four subsectors of the digital economy: (i) *Telecom Companies*; (ii) *Mobile/Internet Contents and Services*; (iii) *SW&IT Services* and (iv) *Application Software.* According to the Value at Risk estimates, we find important differences between markets.

IV. Empirical Application

A. Data Analysis

The data used are the closing prices of daily stock quotes obtained from Yahoo Finance. The two stock indexes where they have been quoted are in NasdaqGS and in the NYSE, with the exception of the Samsung C & T Corporation, which is traded only in Korea. The reference period is January 3, 2000 to July 15, 2014. The sample used consists of 29 firms.[3] These companies have been classified into four homogeneous subsectors: *Telecom*, *Mobile/Internet Contents & Services*, *SW&IT Services* and *Application Software*. The four subsectors comprise the following companies: *Telecom* is composed of AT&T, BT, Orange, Telefónica, Verizon, Vodafone, Vimpelcom and Shaw Communications. *Mobile/Internet Contents & Services* consists of Amazon, Apple, Ebay, Priceline, Yahoo and Netflix. *SW&IT Services* contains Microsoft, Oracle, IBM, Citrix, Ca Technologies, Adobe and SAP. *Application Software* consists of ADP, Autodesk, Cerner, Cognizant, Fiserv, Intuit and Symantec.

Given the high number of companies that integrate the different subsectors, it is helpful to elaborate indexes that allow us to aggregate the information and facilitate the analysis. In this regard, we define a representative portfolio per subsector

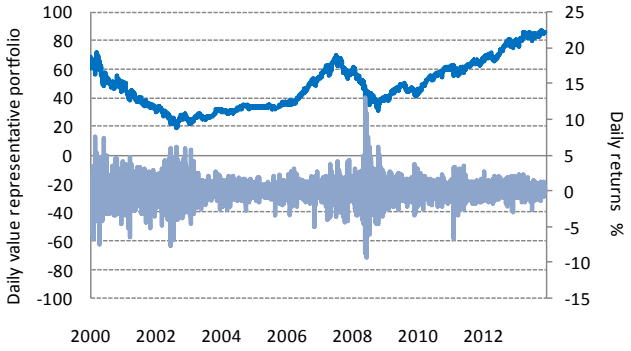


Fig. 2. Telecom

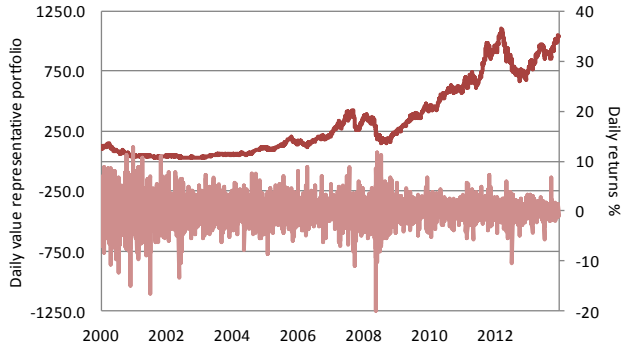


Fig. 3. Mobile/Internet

[2] To estimate the Value at Risk of a portfolio**,** several methodologies have been developed: (i) the parametric approach; (ii) the non-parametric approach and (iii) the semi-parametric method. The parametric approach is the one most used by financial institutions. In Appendix A, we describe carefully how to calculate VaR using this methodology and summarize the backtesting procedure used to evaluate the VaR estimate.

[3] The initial sample consists of 40 companies within the digital economy and the service sector. However, we have eliminated 12 firms because they are young companies listed on the stock exchange after January 3, 2000.

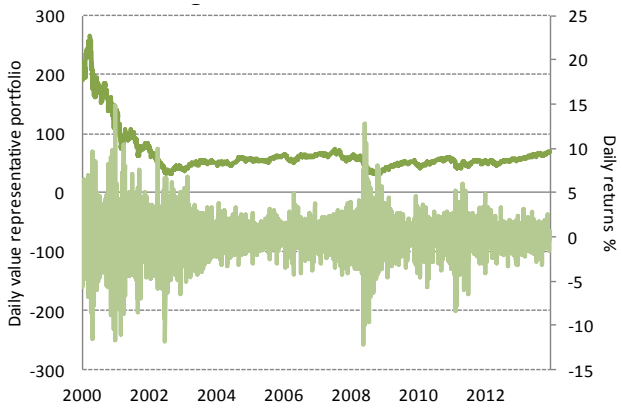


Fig. 4. SW&IT Services

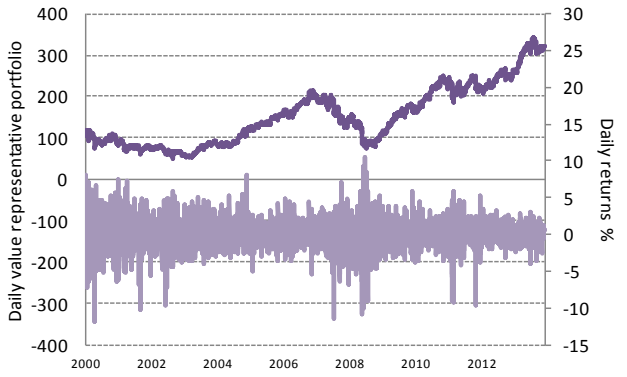


Fig. 5. Application Software

The evolution of the daily value of the portfolio of *SW&IT Services* shows a strong decline from late 2000 to mid-2003. In the remaining period, there were slight variations in the portfolio value. In this subsector, we observed a moderate profile that is typical of a mature market, even a technologically advanced sector. This is indicative of the rate of advances occurring in the field of Digital Economy. In the evolution of the daily value of the *Telecom* portfolio, the range of fluctuation is very small (between 19.3 and 86.9), so the prices are quite stable in this portfolio. The effect of the international financial crisis is hardly felt.

Table 1 contains descriptive statistics of the returns per portfolio. For each portfolio, the unconditional mean of daily return is very close to zero (for example, Telecom (0.01)).

TABLE I
DESCRIPTIVE STATISTICS OF DAILY RETURNS SERIES

	Telecom	**Mobile / Internet**	***SW&IT Services***	***Application Software***
Mean	0.01	0.06	-0.03	0.03
Std. Deviation	1.5	2.51	2.14	1.89
Skewness	-0.02	-0.34**	-0.16**	-0.37**
Kurtosis	5.92**	4.16**	4.66**	4.03**
Minimum	-9.26	-20.04	-12.09	-11.97
Maximum	13.11	14.84	14.84	10.5

Note: ** denotes significant statistics at 1% level.

It means that the fluctuations of positive and negative daily returns of each portfolio tend on average to offset. The unconditional standard deviation is especially high for Mobile/Internet Contents & Services (2.51), followed by SW&IT Services (2.14) and Application Software (1.89). Therefore, the daily returns on these portfolios vary considerably from day to day. In the first two aforementioned subsectors, their portfolios underwent great changes in the daily returns over the period considered. Far from these estimations, we find the Telecom subsector with a standard deviation of (1.5). These results indicate that according to the traditional measure of risk, the most traditional subsector in the digital economy is the safest. *Furthermore, these results show that* the Telecom *market is a mature sector with few* prospects for growth. On the contrary, the Mobile/Internet Contents & Services sector is the riskiest, typical of a young, dynamic and growing sector.

The skewness statistic is negative in all the portfolios, especially with a more negative value in Mobile/Internet Contents & Services (-0.34) and Application Software (-0.37). Thus, the distribution of those returns is skewed to the left. This means that the probability that the daily return is negative is greater than the probability that the daily return is positive. The excess kurtosis is very large and significant at 1% level, implying that the distributions of those returns have much thicker tails than the normal distribution. These results are in line with those obtained by Bollerslev [3], Bali and Theodossiou [4], and Bali et al. [5], among others. All of them find evidence that the empirical distribution of the financial return exhibits a significant excess of kurtosis (fat tails and peakness). This descriptive analysis reveals that Mobile/Internet Contents & Services is the most asymmetric, plus one of the most volatile subsectors.

B. The Value at Risk of the portfolio representative

In this section, we calculate the Value at Risk of a portfolio representative of each subsector. The VaR ha**s** been calculated at one day ahead at 1% probability. The data period is divided into a learning sample from January 3, 2000 to December 31, 2007 and a forecast sample from January 1, 2008 to the end of June 2014.

In Figures 6 to 8, we present the daily VaR estimates for each representative portfolio, hereinafter expected losses. After the fall of the Lehman Brothers, in September 2008, the expected losses increased considerably in all markets, reaching double figures. This increase must be contextualized in a context of global financial and economic crisis which continued until the end of 2009. In 2010, expected losses returned to pre-crisis levels. Two years later, expected losses again soared, although this time with less intensity than in 2008. Although in qualitative terms the expected losses have a similar behaviour in all markets, in quantitative terms there are significant differences between them.

In Figure 8, we present the average of the daily Value at Risk estimate, i.e., the average of the expected losses for each portfolio.

During the analyzed period, the highest expected losses are observed in the Mobile/Internet Contents & Services (-5.3%)

market subsector, while in the Telecom (-3.1%) sector, the expected losses are the lowest.[4] These percentages mean that, for instance, for a portfolio value of 500,000 euros, the expected losses would be 26,500 euros in the *Mobile/Internet Contents & Services* sector and 15,500 euros in the *Telecom* sector.

Thus, the risk of investment in Mobile/Internet Contents & Services companies is higher than Telecom companies.

However, we have not observed significant differences between SW&IT Services and Application Software.

As higher risk implies higher yield, these results are coherent with the fact that the yield of Mobile/Internet Contents & Services companies has been the highest. These companies are the big winners in the **d**igital economy, creating more profitable but riskier investments.

Once we have estimated the Value at Risk per subsector, we proceed to evaluate the accuracy of the VaR estimates and calculate the average of the unexpected losses.

Table 2 shows the percentage of exceptions for the 4 subsectors. These percentages are marked in bold. Below the percentages, we present the statistics used to test the accuracy of the VaR estimates. These statistics are as follows: (i) the unconditional coverage test (LRuc); (ii) statistics for serial independence (LRind); (iii) the conditional coverage test (LRcc) and (iv) the dynamic quantile test (DQ). In all subsectors, the percentage of exceptions is very close to the theoretical level, which is 1%, so it seems that the VaR estimates are accurate in all markets. This result is corroborated by the statistical test.

Table 3 displays the average of the no cover losses which are measured by López´s loss function (see Appendix A). In a comparison between subsectors, we observe that the no cover losses are lower in the *Telecom* sector and *SW&IT Services* sector, which are the most traditional markets among the digital global market. By contrast, in the *Mobile/Internet Contents & Services* market, the no cover losses are the highest, followed by the *Application Software* subsector.[5]

To this point, we can conclude that whatever measure we use to evaluate risk, variance and/or value at risk, we find important differences between the four subsectors of the digital economy. In particular, we find that the riskiest sector is *Mobile/Internet Contents & Services*. In contrast, the *Telecom* sector is by far the safest one.

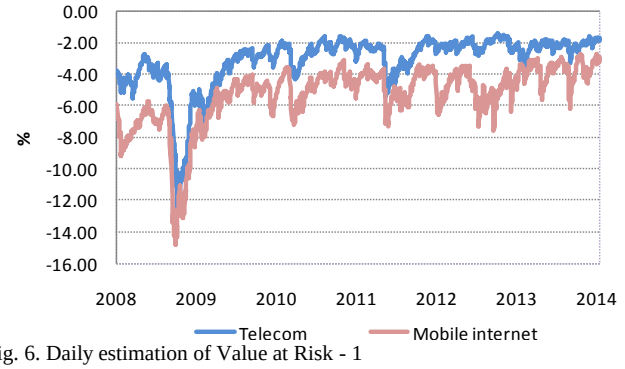


Fig. 6. Daily estimation of Value at Risk - 1

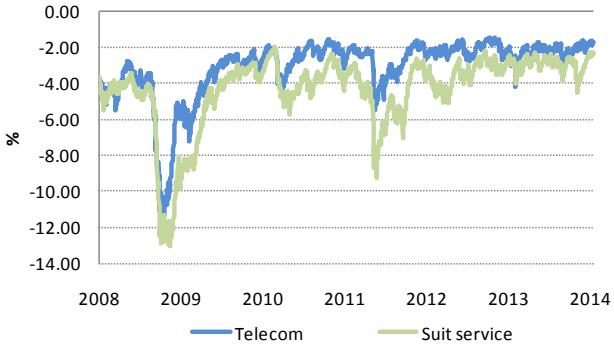


Fig. 7. Daily estimation of Value at Risk - 2

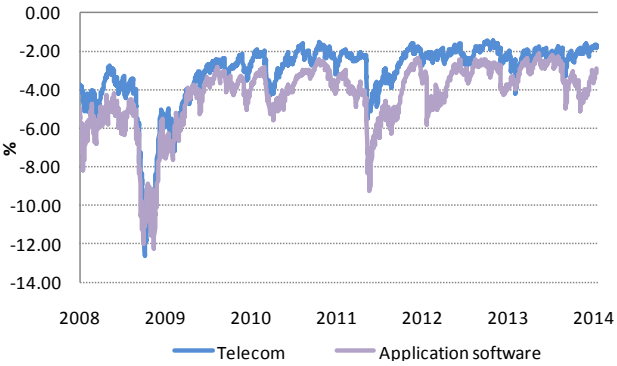


Fig. 8. Daily estimation of Value at Risk - 3

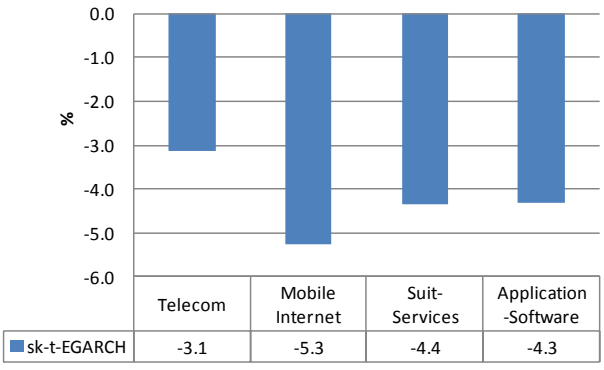


Fig. 9. Average of the daily VaR estimates

[4] To assert the robustness of the result, we have calculated VaR using other volatility models like the standard GARCH [6] and the beta-t-EGARCH model proposed by Harvey and Chakravarty [7]. All of them provide the same results.

[5] Again we find that this result is robust to the volatility model used to forecast VaR, being the standard GARCH [6] and the beta-t-EGARCH model the two alternative models we have considered. These models have been estimated below a symmetric distribution (student-t).

Furthermore, in a comparison between models for each subsector, we find that the Beta-skewness-t-EGARCH model provides by far the lowest no cover losses. From this analysis, we can conclude that the fat-tailed and skewness distributions for conditional volatility outperform symmetric distribution in forecasting VaR. This result is in line with those presented by Xu and Wirjanto [8], Polanski and Stoja [9] and Chen et al. [10].

TABLE II
THE ACCURACY OF VAR ESTIMATES

	Telecom	**Mobile / Internet**	***SW&IT Services***	***Application Software***
	1.10	**0.86**	**1.22**	**1.16**
LR_{UC}	0.10	0.14	0.63	0.04
LR_{IND}	0.17	0.10	0.23	0.11
LRcc	0.27	0.23	0.86	0.15
DQ	0.72	0.10	2.27*	1.28

Note: LRuc denotes the unconditional coverage test; LRcc denotes the conditional coverage test; LRind: denotes the independence test; DQ denotes the Dynamic Quantile test. * indicates that the null hypothesis is rejected at 5% level.

TABLE III
MAGNITUDE OF THE LOSS FUNCTION

	Telecom	**Mobile / Internet**	***SW&IT Services***	***Application Software***
Average of the loss function	**0.02**	0.11	0.04	0.05

Note: Bold figures denote the minimum value of the loss function.

V. CONCLUSION

The digital economy sector is not a homogeneous sector. Four distinct subsectors are observed: *Telecom Companies*; *Mobile/Internet Contents and Services*; *SW&IT Services* and *Application Software*.

The nature of the business models of these subsectors is different, each one presenting very different risk profiles.

According to the traditional measure of risk, which is variance, the riskiest subsector is *Mobile/Internet Contents & Services*, followed by *SW&IT Services and Application Software*. In contrast, the *Telecom* sector is by far the safest one.

Measuring risk through value at risk methodology, which is the most commonly used at present, the results are qualitatively similar. According to this methodology, the highest expected losses are observed in the *Mobile/Internet content & Services* subsector, while in the *Telecom* sector, the expected losses are the lowest. That implies that the risk of investment in *Mobile/Internet Contents&Services* companies is higher than in *Telecom companies*.

As higher risk implies higher yield, these results are coherent with the fact that the yield of Mobile/Internet Contents & Services companies is the highest in the **d**igital economy market.

In addition, we find that no cover losses are lower in the *Telecom* sector and the *SW&IT Service*s sector, which are the most traditional market**s** among the digital global market. In contrast, in the *Mobile/Internet Contents & Services* market, the no cover losses are the highest, followed by the *Application Software* subsector.

APPENDIX A. VALUE AT RISK METHODOLOGY

In this appendix, we present the concept of Value at Risk (VaR) and how to calculate it. In addition, we summarize the backtesting procedure we have used to evaluate the VaR estimates.

According to Jorion [11], the VaR measure is defined as the worst expected loss over a given horizon under normal market conditions at a given level of confidence. The VaR is thus a conditional quantile of the distribution of asset returns. Use F(*r*) to denote the cumulative distribution function, $F(r) = \Pr(r_t < r / \Omega_{t-1})$, conditionally on the information set Ω_{t-1} that is available at time *t-1*. The VaR with a given probability $\alpha \in (0.1)$ denoted by VaR(α) is defined as the quantile of the probability distribution of financial returns: $F(VaR(\alpha)) = \Pr(r_t < VaR(\alpha)) = \alpha$

Under the framework of the parametric techniques [11], the conditional VaR can be calculated as $VaR_{t+1}(\alpha)=\mu_{t+1}+\sigma_{t+1} * k\alpha$, where μ_{t+1} represents the conditional mean, which we assume is zero, σ_{t+1} is the conditional standard deviation and kα denotes the corresponding quantile of the distribution of the standardized returns at a given confidence level 1-α. For instance, if we assume a normal distribution for the financial returns, kα will be the quantile of the standardized normal distribution.

In this paper, we forecast value at risk one day ahead at 1% probability. To forecast the conditional standard deviation of the return portfolio which is required for the VaR estimate, we use the Beta-skewness-t-EGARCH model proposed by Harvey and Sucarrant [12]. This model has been estimated below an asymmetric student-t distribution.

To test the accuracy of the VaR estimate, we use several standard tests: unconditional coverage tests, conditional coverage tests, the Back-Testing criterion and the Dynamic Quantile test.

We have an exception when $r_{t+1} < VaR_{t+1}(\alpha)$; in this case, the exception indicator variable (I_{t+1}) is equal to one (zero in other cases). Kupiec [13] shows that the unconditional coverage test has as a null hypothesis, with a likelihood ratio statistic given by

$$LR_{UC} = 2\left[log\left(\hat{\alpha}^x (1-\hat{\alpha})^{N-x}\right) - log\left(\alpha^x (1-\alpha)^{N-x}\right)\right]$$

which follows an asymptotic $\chi^2(1)$ distribution. The conditional coverage test [14] jointly examines whether the percentage of exceptions is statistically equal to the expected percentage and the serial independence of I_{t+1}. The likelihood ratio statistic of the conditional coverage test is LRcc=LRuc+LRind, which is asymptotically distributed $\chi^2(1)$, and the LRind statistic is the likelihood ratio statistic for the hypothesis of serial independence against first-order Markov dependence. Finally, the dynamic quantile test proposed by Engle and Manganelli [15] examines whether the exception indicator is uncorrelated with any variable that belongs to the information set Ω_{t-1}available when the VaR was calculated. This is a Wald test of the hypothesis that all slopes are zero in a regression of the exception indicator variable on a constant, five lags and the VaR.

Additionally, we evaluate the magnitude of the losses experienced. For this purpose, we have considered the loss function proposed by Lopez [16],[17]. This function reflects the utility function of a regulator. This loss function assigns a quadratic specification when the observed portfolio losses exceed the VaR estimate. Thus, we penalize only when an exception occurs according to the following quadratic specification:

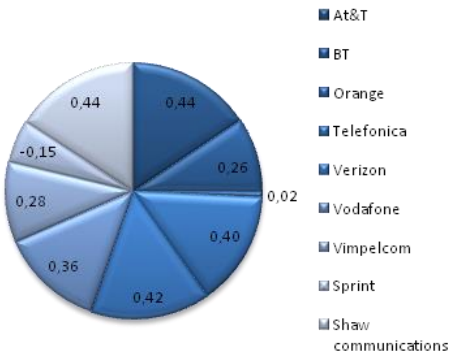
$$RLF_t = \begin{cases} (VaR_t - r_t)^2 & if \quad r_t < VaR_t \\ 0 & otherwise \end{cases}$$

This loss function gives higher scores when failures take place and considers the magnitude of these failures. In addition, the quadratic term ensures that large failures are penalized more than small failures.

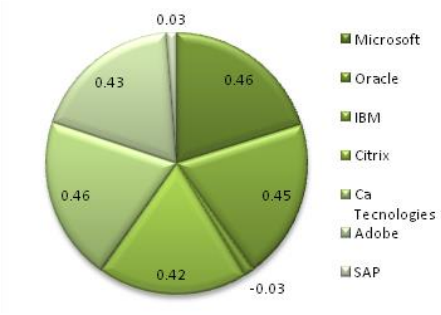
APPENDIX B. THE METHOD OF PRINCIPAL COMPONENTS ANALYSIS

We use the method of Principal Components Analysis (PCA) to calculate linear indexes constructed using the closing daily prices per firms. PCA is a statistical procedure used to reduce the dimensionality of a data set. Intuitively, the technique finds the causes of variability in a data set and sorts them by importance. PCA uses an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components. This transformation is defined in such a way that the first principal component has the largest possible variance (that is, accounts for as much of the variability in the data as possible), and each succeeding component in turn has the highest variance possible under the constraint that it is orthogonal to (i.e., uncorrelated with) the preceding components. The first principal component is used to derive the weights of the index. Figure A1 summarizes the first principal component per subsector. Each graph shows the weight that each firm has in the composition of the first principal component per subsector. Therefore, the first component can be interpreted as a representative portfolio per subsector. In each subsector, the total explained variable is over 50%, with the exception of the Mobile/Internet sector (46%).

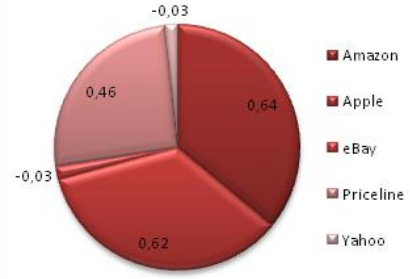
Telecom



Mobile / Internet Contents & Services



Application Software



SW&IT Services

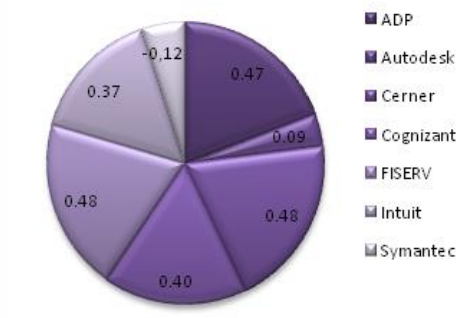


Fig. 10. The first principal component per subsector

REFERENCES

[1] J. Calderón, "La historia de un mundo digital," in *Economía digital. Curso MOOC*. F. Mochón, J. C. Gonzálvez and J. Calderón. Alfaomega, 2014.

[2] Goldman Sachs. Publications by Sector/Industry https://360.gs.com/gs/portal/research/industrylist.

[3] Bollerslev, "A Conditionally heteroskedastic time series model for speculative prices and rates of return." *Review of Economics and Statistics*, vol. 69, pp. 542-547, 1987.

[4] T. Bali and P. Theodossiou, "A conditional-SGT-VaR approach with alternative GARCH models," *Annals of Operations Research, vol.* 151, pp. 241-267, 2007.

[5] T. Bali, M. Hengyong and Y. Tang, "The role of autoregressive conditional skewness and kurtosis in the estimation of conditional VaR." *Journal of Banking & Finance* vol. 32(2), pp. 269-282, 2008.

[6] T. Bollerslev, "Generalized Autoregressive Conditional Heteroscedasticity." *Journal of Econometrics*, vol. 21, pp. 307-327, 1986.

[7] A.C. Harvey and T. Chakravarty, Cambridge Working paper in Economics, CWPE 0840, 2008.

[8] D. Xu, T. Wirjanto, "An empirical characteristic function approach to VaR under a mixture-of-normal distribution." *The Journal of Derivatives*, vol. 18(1), pp. 39-58, 2010.

[9] A. Polanski and E. Stoja, "Incorporating higher moments into Value-at-Risk forecasting." *Journal of Forecasting*, vol. 29(6), pp. 523-535, 2010.

[10] C. Chen, R. Gerlach, E. Lin and W. Lee, "Bayesian forecasting for financial risk management, pre and post the global financial crisis." *Journal of Forecasting*, vol. 31(8), pp. 661-687, 2012.
[11] P. Jorion, *Value at Risk: The New Benchmark for Managing Financial Risk*. McGraw-Hill, 2001.
[12] A. Harvey and G. Sucarrant, "EGARCH models with fat tail, skewness and leverage." *Computational Statistics and Data Analysis*, vol. 76, pp. 320-338, 2014.
[13] P. Kupiec, "Techniques for Verifying the Accuracy of Risk Measurement Models." *Journal of Derivatives*, vol. 2, pp. 73-84, 1995.
[14] P. Christoffersen, "Evaluating interval forecasting." *International Economic Review*, vol. 39, pp. 841-862, 1998.
[15] R. Engle and S. Manganelli, "CAViaR: Conditional Autoregressive Value at Risk by Regression Quantiles." *Journal of Business & Economic Statistics*, vol. 22 (4), pp. 367-381, 2004.
[16] J.A. Lopez, "Testing your Risk Tests." *Financial Survey*, pp. 18-20, May-June 1998.
[17] J.A. Lopez, "Methods for Evaluating Value-at-Risk Estimates." *Economic Review Federal Reserve Bank of San Francisco*, vol. 2, pp. 3-17, 1999.

Sonia Benito Professor, Department of Economics, UNED, (Spain). Ph.D. in Economics from the Universidad Complutense (Spain). She has pursued her research in Empirical Finance: interest rate common factors, financial returns volatility and finance risk management (Value at Risk). She is the author of several articles, published both in Spanish and international magazines, particularly in the field of finance. She also took part in five projects (CICYT) funded by the Spanish Ministry of Education between 2003 and 2015. She has participated in more than 15 national and international congresses, conferences and seminars, mainly as a speaker.

Rebeca de Juan, Professor, Department of Economics, UNED (Spain). Ph.D. in Economics from the UNED (Spain) and Master in Economic Analysis and Public Finance from the Instituto de Estudios Fiscales. She has been a researcher in the Program for Economic Research at the Fundación Empresa Pública (1994-2002). The research is focused on industrial organization, the banking sector, happiness and digital economy. This research has led to participation in various projects (CICYT) and in private research projects. She has received several special awards. She has published several articles and books.

Ricardo Gómez. Holds an MSc in Industrial Engineering at Universidad Politecnica de Madrid and an MBA from IESE Business School. Currently is Chief Operations Officer at Zed Group where he has formerly held several positions: Organization&Project Director and Corporate Development Director. Prior to joining Zed, held several finance and management positions at Grupo Ferrovial, Cintra and Accenture.

Francisco Mochón. PhD in economics from Indiana University and the Autonomous University of Madrid. Currently is full Professor of Economic Analysis at the UNED. He has been Advisor to the Ministry of Economy and Finance of Spain, and Chief Financial Officer (CFO) of Telefónica of Spain. Currently is a member of the advisory committee of the Futures Market of Olive Oil (MFAO). He has published numerous research articles and is the author of more than fifty books on economics, finance and business. Currently his research interest is the Digital Economy.

The Digital Economy: Social Interaction Technologies – an Overview

Teófilo Redondo

ZED Wordwide, Department of Research and Innovation, Madrid, Spain

Abstract **— Social interaction technologies (SIT) is a very broad field that encompasses a large list of topics: interactive and networked computing, mobile social services and the Social Web, social software and social media, marketing and advertising, various aspects and uses of blogs and podcasting, corporate value and web-based collaboration, e-government and online democracy, virtual volunteering, different aspects and uses of folksonomies, tagging and the social semantic cloud of tags, blog-based knowledge management systems, systems of online learning, with their ePortfolios, blogs and wikis in education and journalism, legal issues and social interaction technology, dataveillance and online fraud, neogeography, social software usability, social software in libraries and nonprofit organizations, and broadband visual communication technology for enhancing social interaction. The fact is that the daily activities of many businesses are being *socialized*, as is the case with Yammer (https://www.yammer.com/), the social enterprise social network. The *leitmotivs* of social software are: create, connect, contribute, and collaborate.**

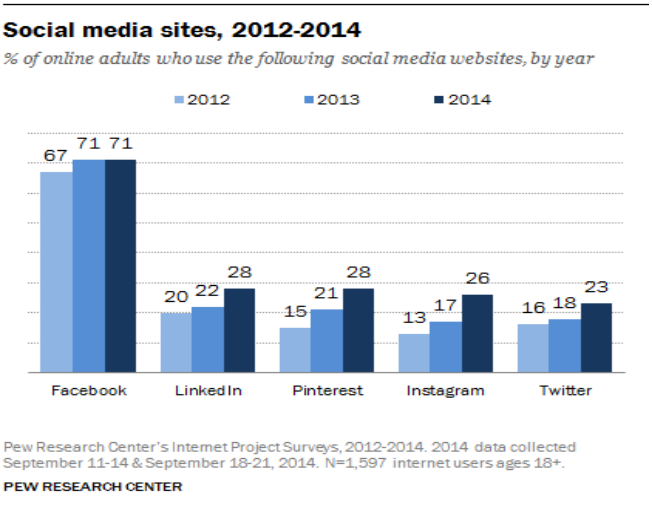
Keywords **— blogs, folksonomies, online learning, social interaction technologies, social media, social web, wikis**

I. Introduction and Background

In recent years, we have been bearing witness to an exponential growth in capabilities to electronically collect, process, store, retrieve and disseminate information and create new knowledge. This has been the case with Internet-based collaboration tools and platforms reaching end-users in unprecedented ways: online social networking, blogs, wikis, podcasts, web feeds, folksonomies, social bookmarking, photo and video sharing, discussion forums, virtual worlds, all intended to advance interaction, collaboration, and sharing online. Social Computing is the generic term used to refer to any type of computing where software serves as an intermediary for a social relation. In social computing the user takes an active role in the process, often creating content or modifying previous content, and the computing experience has extended from the individual to the social.

Social interaction technologies (SIT) and collaboration software touch many fields and they have impacted on many fields, that is to say that Web 2.0 communication converge both socially and technologically. Web science must be interdisciplinary, since it brings together experts from computer science, software engineering, management information systems, business and economics, knowledge management systems, marketing, public relations and advertising, law, journalism and media, communication, psychology, anthropology, social work, design, library and information science, and education. The new emphasis is on user-generated content, creativity, and community-based knowledge building are characteristic of Web 2.0. The term "Web 2.0" suggests a fundamental technological improvement by assigning a version number like typical IT products. Common characterizations of Web 2.0 are mostly based on seven paradigms defined by Tim O'Reilly, in his now famous blog entry *What is Web 2.0* [1]: "The Web as a Platform," "Harnessing Collective Intelligence," "Data is the Next Intel Inside," "End of Software Release Cycle," "Lightweight Programming Models," "Software above the Level of a Single Device," and "Rich User Experiences."

In this "Web 2.0" world, web users have begun publishing their own content on a large scale and started using social software to store and share documents, such as photos, videos or bookmarks. The current trends indicate that a large number of US adult online consumers make daily use of social networking sites, publishing blogs/webpages, uploading visual content, commenting on blogs, posting reviews, or simply consuming user-generated content.



Fir.1. Social Media Sites (Source: PewReseach Internet Project: http://www.pewinternet.org/2015/01/09/social-media-update-2014/)

The theoretical foundation of the social web can be traced back to J.C.R. Licklider with his seminal article "Man-

DOI: 10.9781/ijimai.2015.322

Computer Symbiosis," [2] published in 1960, the first of three articles that attempted to redefine the human-computer interaction. Licklider outlined a vision for interactive, networked computing and, ultimately, the Internet that we experience today.

Licklider's career did not originate in computing; his studies began in "physiological psychology," the field known today as neuroscience. Licklider investigated the brain's ability to understand speech in the presence of signal distortion [3]. These early studies helped Licklider understand the workings of the human brain and prepared him to foresee the potential for improved human-computer interactions, both individually and collectively, the very basis of the online communities that conform the Social Web.

The term "Social Web" is often used in everyday language as well as in academic literature as a synonym for "virtual" and "online communities". An online community is seen as a social group that interacts through a web platform over an extended period of time. An online community can be characterized by four elements:

- group of people with shared objectives (e.g., interests, goals)
- interaction over an extended period of time
- closeness due to bonds and relationships
- shared space for interactions governed by certain rules (for example, role definitions)

The Social Web refers to an aggregation of social interaction and collaboration technologies and can be viewed as a concept and a platform that utilizes social software to support some human needs. The Social Web encompasses numerous Internet applications, such as social networking sites, blogs, podcasts, wikis, massively multiplayer online role-playing games, photo and video sharing, online stores and auction houses, simulated 3-D virtual worlds, and wiki collaborations. Various attempts to provide a definition for the Social Web have resulted in three different approaches: technical, social, and economic. The technical approach focuses on the Internet as a medium or platform for a community. The sociological point of view stresses the forming and functioning of communities, whereas the economic perspective examines potential gains and intended profits [4].

MIT's Media Lab and Intel Corporation each developed two early mobile social web applications. *Social Serendipity* was MIT's Bluetooth-based social service meant to harness the power of mobile technology and social information. *Social Serendipity* facilitated social interaction among geographically proximate users by matching user profiles and then exchanging profile information with similar matches. Intel's *Jabberwocky* sought to monitor and broadcast a user's movement to identify "familiar strangers" and encourage a sense of urban community. Both of these technologies relied on the mobility of the devices to evaluate locational information to facilitate social connections among users.

The Social Web is realized through s*ocial software*, which is a combination of various social tools within a growing ecosystem of online data and services, all joined together (aggregated) using common protocols, and Application Programming Interface (API) methods. Social software is at the center of the so-called API economy, a collective term referring to the economic effects enabled by companies, governments, or individuals providing direct programmable access to their systems and processes through exposing specific APIs for creating larger applications and solutions.

II. SOCIAL SOFTWARE

Several tools are associated with social software:

1. tools allow people to participate by creating, publishing and distributing content, such as video, pictures, music and texts through the Internet.
2. social software allows people with similar interests to find one another and connect through social networking sites, such as Facebook.
3. people can coordinate their activities and collaborate through raising petitions and funds, and planning and conducting mobile campaigns and communities programs.
4. people can create reliable, robust, and complex products such as open source software applications such as Linux (the largest example of community development).

There are three characteristics commonly attributed to *social software*:

- conversational interaction between individuals or groups,
- social feedback that allows a group to rate the contributions of others.
- social networks to explicitly create and manage a digital expression of people's personal relationships.

Social software serves many purposes:

- *Delivery* of communication between groups
- *Enabling* communication between many people
- *Providing* gathering and sharing resources
- *Delivery* of collaborative collecting and indexing of information
- *Providing* new tools for knowledge aggregation and creation of new knowledge
- *Delivery* to different platforms depending on the creator, recipient, and context.

In summary, above all Social Software is about group interaction. For instance, mobile social networks allow users to connect with each other, share information, and create technologically enabled mobile communities. With the introduction of the iPhone in 2007, the public dream of mobile computing was realized. Mobile communication is becoming ubiquitous in many parts of the world today with over 4 billion mobile phone users worldwide.

Although mobile phones may lead to the atomization and privatization among users by discouraging face-to-face communication, the instant accessibility to whatever social app validates the social effects of mobile phone use. As mobile technology advances, new services for mobile phones have

been developed, which allow people to create, develop, and strengthen social interaction.

Next we will explore a little bit further the most common Social Interaction Tools.

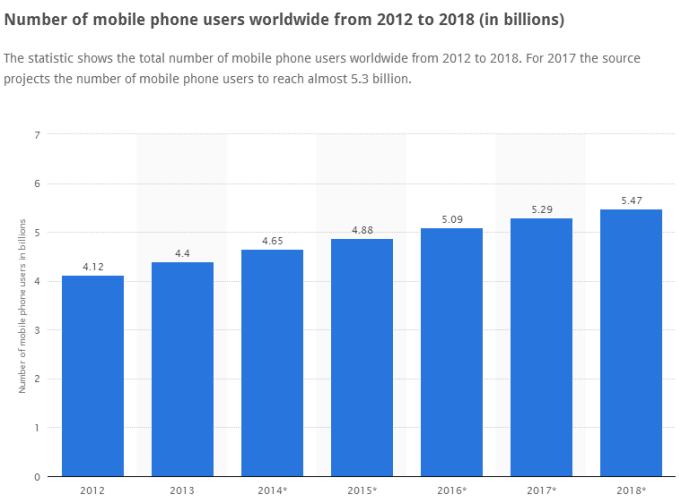


Fig. 2. Number of mobile phone (Source: Statista – The Statistics Portal: http://www.statista.com/statistics/274774/forecast-of-mobile-phone-users-worldwide/)

III. Social Interaction Tools

A. Discussion forums

The origin of Internet forums can be traced back to Usenet and its newsgroups, and the most recent form are the comments section at the end of articles in newspapers/magazines web sites. A forum provides an online exchange of information between people about a particular topic. It can be a place for questions and answers, or a comment-centric site and may be monitored to keep the content appropriate. Forums can be either text-only or media-rich (using images or videos to illustrate a point), and sometimes can be like a mini-portal on the topic. Forums can be entirely anonymous or require registration with username and password.

There are many types of forums: educational, professional, political, … for instance, forums have been institutionalized as an integral part of the political communication system.

B. Blogs and Wikis

One of the main results of the Web 2.0 are new modes of online communication and self-expression. The basic form of a blog is generally accepted to be brief posts, collected on one web page, which are chronologically ordered rather than by topic or argument. Blogs have a commentary concentrated style, which may also include links, pictures, video, or any other media forms, as well as reader comments. There are two main categories: (a) *filter blogs* tend to be focused on external events with all kinds of links, and (b) *journal blogs*, a blog created and maintained by an individual where authors write about events in their own lives. Blogs can be viewed as both a technology and a communication channel.

Blogs are sometimes used by organizations as educational or knowledge management tools. Organizations are using blogs for internal information sharing and knowledge management applications. Information shared within corporate blogs may include: industry or company news, strategy brainstorming, activities within a specific department, and the sharing of customer related information. But a number of concerns have arisen, mainly a preoccupation with regards to productivity. These concerns include: a waste of work time, loss of productivity, posting of inappropriate content, reluctance of employees to share knowledge, and increasing bandwidth requirements.

A number of sites provide all the necessary tools to get started in blogging: *Wordpress*, *Typepad* or *Blogger* are just a few of them.

The term Wiki refers to both technology and a concept of how one can create and edit online content. Wikis are editable websites that enable users to build content and collaborate. Wikis allow users to browse through Wiki pages, edit and modify existing pages, or develop new content in a collaborative way. A Wiki is both a website and a database for keeping track of all versions of the site as modified by the users. *Wikipedia* [5] is probably the most well-known of wikis, but there are many, some as singular as the *Intellipedia* [6], a wiki for sharing restricted information among intelligence services in the United States.

Ward Cunningham conceived the first Wiki in 1994 as a series of index cards stacked upon one another, each able to be changed or altered individually without affecting the other. The format allowed for each "card" to be a new alteration to the web page and to the central database.

Wikis were first adopted by businesses as collaboration software to allow a large number of people to work on a single idea in real time, where people separated geographically could all give input into a single database using a common format. Depending on how the moderator of the Wiki sets the parameters, users can edit, add or subtract information, remain anonymous or have to login to use the system.

C. Online social networking

Internet has made possible for people to connect with each other beyond geographical frontiers. Social networks encompass interactions between different people, members of a community or members across different communities. Each person in this social network is represented as node or vertex and the communications represent the links or edges among these nodes. A social network comprises several focused groups or communities that can be treated as subgraphs, in order to study the structural and temporal characteristics of social networks. To separate social networks from other types of social interactions we focus here on friendship networks ("friend" in Facebook, or "follower" in Twitter, show this connectivity). In these social friendship networking sites users explicitly provide trust ratings to other members.

Social networking sites allow individuals to create online profiles providing information about themselves and their interests, create lists of users with whom they wish to share information, and to view information published within the

network by their friends. Social networks allow us to share experiences, thoughts, opinions, and ideas.

These social interactions also led researchers to hypothesize "Small World Phenomenon" (also known as "Small World Effect"), everyone in this world can be contacted via a short chain of social acquaintances. The social psychologist Stanley Milgram in a famous experiment in 1967 set out to find the length of this short chain, and resulted in the famous concept, "six degrees of separation". This "connectedness" aspect of social interactions between people have been applied to fields as varied as genealogy studies, mathematics, economy, team sports and even corporate dynamics.

Without providing an exhaustive list of social networks, each one highlighting a special means, we can mention Facebook, Twitter, Instagram, Snapchat, Tumblr, or more specific social networks: business (*LinkedIn*, *Xing*); enterprise (*Yammer*); academic (*ResearchGate*, *Gaudeamus*, *Academia.edu*, *Mendeley*).

D. Virtual worlds

Apart from various social media (Facebook, Blogger, YouTube) Web 2.0 technologies have generated online virtual reality environments (Second Life, World of Warcraft, Sims) that have influenced today's students in many ways. There are good opportunities for immersive experiences within user-constructed environments, communities and quests. Hands-on learning (even virtual hands-on) provide an intense engagement of immersive cognitive responses.

One particular case is Massively Multiplayer Online Role-Playing Games (MMORPGs), where participants move from loose to strong associations forming social networks via structured guidelines and interaction patterns. These virtual world inhabitants create communication conduits, collaborate to attain goals and solve problems, or entertain themselves, to chart the various associations ranging from casual conversations to groups in which role specialization is critical to community success. The basis for gamification lies in using rewards for accepted behavior, then creating a socialization continuum that stimulates players to interact with one another.

The vitality of MMORPGs and MMOGs (Massively Multiplayer Online Games) assure that more MMORPGs will be on the way [8].

E. Folksonomies

Folksonomy is a portmanteau of the words folk and taxonomy. Folksonomies are a type of annotation usually referred to as social tagging, with the purpose of knowledge representation and knowledge management. Social tagging is a multidisciplinary linking knowledge representation and classification that creates an open domain network. Many recent tools and techniques focus on exploring aspects of the connection between social tagging and the underlying community, in particular the role of tagging as a means of shared informal annotation. These tags are based on user motivation and function.

Folksonomies are a relatively novel way of indexing documents and locating information based on user generated keywords. This type of grass-roots community classification (similar to other social networking approaches, such as blogs and wikis) is a good example of collective intelligence. While taxonomies are hierarchical classifications defined by formal methods that do not necessarily include user-generated tags, folksonomies structure content via user tags and the vocabulary is not preassigned. Hierarchical taxonomies attempt to organize information and give context to data through a branching structure while folksonomies allow for a multiplicity of contexts.

A folksonomy can be seen as an indexing method open for users to apply freely chosen index terms. Peter Merholz [9] entitles this method "metadata for the masses". The term "folksonomy", was introduced in 2005 by Thomas Vander Wal [10], who defines folksonomy as the outcome of individual free tagging of online content and resources in a social environment for one's own retrieval, collaboratively assigning keywords to resources or items, and sometimes have been used synonymously with the terms social classification, social indexing, or social tagging.

Large-scale social bookmarking sites (such as *Del.icio.us* or *Reddit*) have been among the earliest adopters of using folksonomies to organize information. These sites are effective tools for storing, finding, and sharing Internet-based resources, a form of social knowledge management. Much of the success of social bookmarking is attributed to its loosely structured approach to organizing data and the ease with which consumers can learn and integrate tags into a folksonomy.

F. Podcasts and webcasts

Podcasting, a portmanteau word created out of the brand name iPod and the term broadcasting is a distinctive area within social interaction technology. Content is often listened to or viewed within the world of a personal audio/video device, this more so with the ubiquitous mobile device. The upload or the download is the interaction. Podcasts are used for entertainment, education, instruction, profit and just to enjoy some time.

The first podcast took place in 2003 by automatically streaming a single audio file half way around the world. Months later Apple Computer, Inc. proved that its personal listening device, the iPod, could synchronize with a new program called iTunes and download files using the same technology. The broadcast media have begun to use podcasting as a method of time shifting programming, to allow its audience to listen to content claiming the audience consume "what they want, when they want". Educators are using podcasting for reaching out to students. Businesses are using podcasting as a marketing tool. The commercial future of podcasting appears to be in the area of advertising and broadcasting.

The user listens to or views the file, deletes it, and waits for the next episode. Because a podcast audio or video file is digital and in a format common to many devices (normally MP3), that file can go viral being reposted, edited, linked to through social sites (such as YouTube, or Facebook), or moved around through email or by some other social interaction means.

Webcasts in the form of broadband visual communication (BVC) technologies (such as videoconferencing and video sharing) allow for the exchange of rich simultaneous or pre-recorded visual and audio data over broadband networks. BVC involves both simple and complex social and technical interactions. The complexities arise as the interaction grows from communication between two individuals in the same location to communication between multiple individuals in multiple locations, working for multiple organizations located in different communities.

G. Photo and video sharing

Photo sharing can be said to be one of the first social engaging uses of the early Web 2.0 days, even before the term was coined. At a time when most cameras were not digital, users started uploading (publishing their digital photos online) and providing links for friends to share and comment, using websites like *Picasa*, *Flickr*, or *Instagram*.

Similarly, video sharing sites allow users to upload and share their video clips with their friends or connections (private) or with public at large (public). The best known examples are *YouTube*, *Vimeo*, and *Dailymotion*, as well as *Netflix*, *Hulu*, or *Vine*. Apart from traditional video some of these sites provide webcasting capabilities too. For instance, YouTube personal channels can be used for streaming content, which is particularly useful in the case of online learning.

H. Geotagging

Neogeography refers to geography in the Web 2.0 style, a collaborative technology from the public rather than from those in the profession, that is, a group of people (many unknown to one another) who volunteer collectively to contribute data about a topic, in this case, mapping. The practice of neogeography shares the characteristics of other social interactive technologies. Volunteer-supplied geographic tags may assume informational value beyond entertainment. Neogeography-related websites provide different ways for people to contribute photographs, locations, tags, and comments.

Neogeography might be considered a subset of cybercartography or interactive, web-based spatially referenced data, and can trace its origins to the Geospatial Web or the GeoWeb. Any sort of data that conveys place can qualify as geographical data, including for example, zip codes, area codes, images of a place, census data or place names, using a specific XML format for geographic data known as GML (Geography Markup Language).

Geosocial networking is the result of combining geotagging with social networking by including geographic services and capabilities to enable additional social dynamics, and this is the case by inserting location coordinates assigned to pictures, and then adding those pictures to maps with many applications: *Flickr* or *Instagram* for photographs, or *Panoramio* for Google Earth / Google Maps. Users of geosocial applications like *Yelp*, *Facebook Places* and *Foursquare* (*Swarm*) share their locations as well as ranked recommendations for locations or 'venues'.

IV. Some Special Use Cases

A. Advertising / marketing

Marketing is historically considered an activity that businesses performs to direct the flow of goods and services from producers to consumers. The growth of the Internet and the development of social software have changed this top-down process in the age of citizen marketing, that is, consumers voluntarily posting product information based on their knowledge and experience. Citizen marketers were envisioned by futurist Alvin Toffler [11] who coined the term "prosumers," blending the words producer and consumer, consumers who educated themselves and became involved in the design and manufacture of products. Product here refers to goods, services, brands, companies, organizations, or people, such as political candidates. The product information may take the form of opinions, reviews, videos, to be found on forums, blogs, ratings or opinion sites, social networking sites, video sharing sites, or even on mainstream marketers' websites as consumer reviews or discussion boards.

Consumers seek product information provided by citizen marketers, who are eager to share their experiences and their knowledge of a product. Citizen marketers are not on the company payroll and are not trying to sell anything. They volunteer their time as writers, animators, designers, and videographers to express their opinions about products.

B. Social capital

The social capital framework is applied to illustrate how Web 2.0 tools and techniques can support effective information and knowledge management in organizations. Managing social capital for effective knowledge sharing is a complex process, and Web 2.0 helps by creating a new culture of voluntary, contributive, and collaborative participation.

Research on social capital has been carried out in different disciplines and at different levels depending on the chosen perspective (e.g., Putnam [12]; Fukuyama [13]). Social capital includes the individual and the social aspects and is defined as the sum of actual and potential resources embedded within, available through, and derived from the network of relationships possessed by an individual or social unit. Social capital encompasses both the network and the assets that may be mobilized through that network.

Social capital is also often described in three dimensions: a structural dimension (network character), a relational dimension (trust and social identity), and a content dimension (communication to facilitate social capital). People generate economic, emotional, spiritual, and social value by engaging in social relationships. Social capital is the glue holding communities together with the power of cooperative actions. Social capital is dictated by how networks of individuals in a community create conditions where people are inclined to do things for one another.

C. Virtual Teams

Social interaction technologies have made it possible for teams to exist in a virtual reality. Team members can create, maintain, transmit and influence their competitiveness and

effectiveness. The key requirements for the functioning of successful virtual teams and online culture are building trust, consolidating authentic communication flows and thinking critically.

Traditionally, a team is viewed as a group of people who bring balanced competencies to shared purposes, approaches and performance targets. There is usually synergy between the individual members of the team, which means that when the individual efforts and actions are harmonized, something different and unique is created that could not be produced by any single individual of the team. The concept of a team is expressed through seven vital elements:

1. size linked to the scope of the task
2. members' skills are balanced (basis for interdependence)
3. mutual accountability (members are synergistic and trust each other)
4. synergies of purpose
5. approach
6. performance targets
7. distance between members

D. Online Learning

Online behavior, distributed collaboration, and social interaction are already having a transformative effect on education, triggering changes in how teachers and students communicate and learn. Learners can engage in creative authorship by producing and manipulating digital content and making it available for consumption and critique by classmates, teachers, and a wider audience on the web.

Informal education (or learning) sits outside the traditional educational context and is voluntary, self-directed, lifelong, and motivated mainly by intrinsic interests, curiosity, exploration, and social interaction, and typically lacks the presence of an instructor. Informal learning often is self-paced and visual- or object-oriented. It provides an experiential base and motivation for further activity and learning.

The benefits and challenges of hybrid courses, which blend face-to-face instruction with online learning, leverage opportunities provided by the introduction of web-based social interaction technologies. Hybrid courses continue to evolve to meet the needs of students, instructors, and institutions of higher learning.

Today, more classrooms are equipped with various types of technology including Internet access, integrated projectors for computers, digital board, or audio and video devices. Online education management systems (*Blackboard*, *Sakai*, or *Moodle*), sometimes called learning management systems (LMS), are becoming more commonplace and are enabling communication, learning materials, assignments, and grading to occur online.

Early online learning environments were not engaging, and limited in supporting the interaction, coordination and cooperation between students and instructors, with low levels of confidence while learning at a distance, and low satisfaction levels resulted as a consequence. At earlier times dropout rates were relatively high. In online learning the motivation, and the sense of shared social experience are greatly constrained. New social interaction technologies can improve the social experience and social support of online learning. The members experience the online environment as a social place for learning and not in isolation. Opinion and preference of online versus classroom has turned around and now online instruction is the preferred way when a choice is offered.

TOTAL AND ONLINE ENROLLMENT IN DEGREE-GRANTING POSTSECONDARY INSTITUTIONS – FALL 2002 THROUGH FALL 2012

	Total Enrollment	Annual Growth Rate Total Enrollment	Students Taking at Least One Online Course	Online Enrollment Increase over Previous Year	Annual Growth Rate Online Enrollment	Online Enrollment as a Percent of Total Enrollment
Fall 2002	16,611,710	NA	1,602,970	NA	NA	9.6%
Fall 2003	16,911,481	1.8%	1,971,397	368,427	23.0%	11.7%
Fall 2004	17,272,043	2.1%	2,329,783	358,386	18.2%	13.5%
Fall 2005	17,487,481	1.2%	3,180,050	850,267	36.5%	18.2%
Fall 2006	17,758,872	1.6%	3,488,381	308,331	9.7%	19.6%
Fall 2007	18,248,133	2.8%	3,938,111	449,730	12.9%	21.6%
Fall 2008	19,102,811	4.7%	4,606,353	668,242	16.9%	24.1%
Fall 2009	20,427,711	6.9%	5,579,022	972,669	21.1%	27.3%
Fall 2010	21,016,126	2.9%	6,142,280	563,258	10.1%	29.2%
Fall 2011	20,994,113	-0.1%	6,714,792	572,512	9.3%	32.0%
Fall 2012	21,253,086	1.2%	7,126,549	411,757	6.1%	33.5%

Fig. 3. Online Enrollment (Source: Edudemic – Connecting education and technology: http://www.edudemic.com/2013-survey-online-learning/)

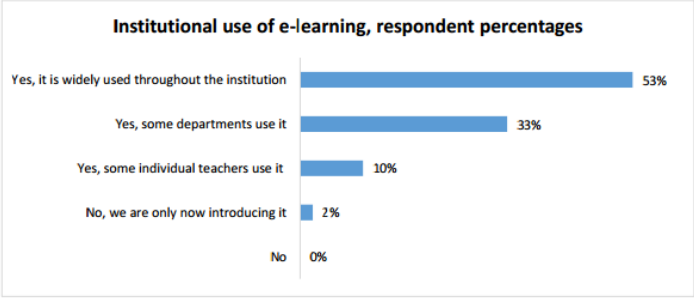


Fig. 4. (Source: European University Association : http://www.eua.be/Libraries/Publication/e-learning_survey.sflb.ashx)

Previous Enrollment in Online Study	Percent of Respondents All	Undergraduate	Graduate
Yes, I took individual online courses	45%	41%	45%
No	43	51	37
Yes, I completed another fully online program	17	11	25

Compare Online to Classroom	Percent of Respondents All	Undergraduate	Graduate
Better	47%	50%	43%
About the same	43	41	48
Not as good	10	9	9

Fig. 5. (Source: Online College Students 2014: http://www.learninghouse.com/wp-content/uploads/2014/06/2014-Online-College-Students-Final.pdf)

Some examples of applying social interaction tools to education follow:

a) Wikis in education

In recent years, the field of education has discovered the educational value of the instructional use of wiki-based classroom technologies. Certain kinds of wiki-related activities correspond to certain levels of classroom interactions: social interaction, general discussion, topic-focused discussion, and collaborative/cooperative activities. Students use Wikis for

collaborative writing exercises, or completing group assignments. [14]

b) Videogames

Squeak Etoys is a free software program and media-rich authoring system with a user-friendly visual interface. The software is designed to help six to twelve year-old children learn through interaction and collaboration. Etoys environment was created to increase the capacity for creative learning, exploration, interaction, and collaboration. Children interact, work together on projects, and engage in computer simulations and games while learning mathematics, physics, chemistry, and geometry. A similar objective is pursued by *Pocket Code*, strongly inspired by MIT's *Scratch*, the free programming language and online community that helps creating interactive stories, games, and animations.

c) Webcasting

Webcasting refers to the delivery of audio and video content over the web. The web is used as a delivery medium for informational, instructional, marketing, and entertainment purposes. Webcasting incorporates social media elements that can assist in the development of a science-oriented educational website.

E. Public libraries

Libraries have started employing social software applications (such as blogs, tagging, social networking, and wikis) to engage readers, encourage user-contributed content, and connect with user populations. User-centered philosophies are at the heart of libraries' service and have been in practice long before the emergence of Web 2.0. However, libraries have seen a radical shift as they are now faced with web-users' expectations. These expectations may not be met with less interactive computer technologies, such as library online public access catalogs (OPACs).

V. Current and Future Trend

A. Semantic Web

Social tagging has become an essential element for Web 2.0 and the emerging Semantic Web applications. Web 2.0 sites express their structure, features, and relations in different ways. The model, termed the Social Semantic Cloud of Tags (SCOT), allows for the exchange of semantic tag metadata and the reuse of tags in various social software applications.

The initial purpose of tagging is to help users organize and manage their own resources, and collective tagging of common resources can be used to organize information via informal distributed groups of users. The power of semantic social tagging lies in the aggregation of information, which involves social cohesion by reinforcing social connections and providing social search mechanisms. A community built around tagging activities can be considered a social network with an insight into relations between topics and users. Semantic Web techniques and approaches help social tagging systems to eliminate tagging ambiguities.

B. eGovernment and privacy in social media

Globalization has brought a special emphasis on knowledge creation and transfer as the primary driver of economic growth and competitiveness with information technologies playing an ever-increasing role. The economic, social and political landscape in which development is taking place has changed completely.

E-government initiatives are aimed at modernizing governmental agencies in their dealings with the public and extending services into online environments. These initiatives have begun in various countries, which have allowed citizens easy access to public services and lobbying opportunities at policy level decision-making.

A number of legal issues around privacy preservation may arise from the increasing use of social interaction technologies: prospective employers searching the Internet to discover information from candidates' blogs, personal web pages, or social networking profiles; or employees being fired because of blog comments. These situations present challenges to legal systems which historically have been slow to adapt to new technologies. As a result, many of these legal issues remain unsettled.

C. Social enterprise

Social software is assuming a significant role in business, and has been utilized recently on a growing scale by companies in customer relationship management (CRM). A firm needs to identify the optimal level of social software deployment when planning to maximize its transactional benefits through the management of a customer knowledge base. The optimal level of social software depends on a range of factors: the initial volume of knowledge base, transaction benefits, and the estimates of the positive and negative effects of social software use.

Only recently have companies started to apply social software for managing customer knowledge, maintaining good customer relationships, and enhancing customer satisfaction, sometimes even reaching to customers in a very personalized manner. Although social software is gradually assuming a more essential role in e-business, it is still unclear at what level firms should implement it. Social software dynamically influences customer knowledge bases with direct and indirect effects of social software implementation on businesses. Customers who are dissatisfied with their shopping experiences may impact the current knowledge base affecting the transactions of future potential customers.

D. Social Web of Things

The relationship between Social Networks and Internet of Things (IoT) was introduced as "Social Web of Things". This idea is a redefinition of the IoT paradigm, in which things leverage social networks and specifically social standards to communicate, assigning a specific social identity to things ("smart objects") at the same level than people. Social Networks can help in that sense in "elevating" the semantics of IoT interactions to the user level and thus fostering the adoption of connected objects such as wearables, home

automation, connected cars, to interconnect different devices with users.

"Smart social objects" consists in creating a network of "trusted" friends between humans and objects. Objects can post information to the social network, show their availability and discover new "friends", interacting with other objects or humans. In that sense the social component adds a user-friendly interaction (dialogue) paradigm for people to interact with their surrounding "Smart Social Space" environment.

Smart Social Spaces could be public, such as a local business, or private, such as a smart office or home, in which appliances and sensors communicate with one another and post their behaviors on the social wall. Users could receive (multimedia) notifications or alerts about sensors and are able to send commands remotely, for example, to their home security cam.

Some current examples of the Social Web of Things are: *Toyota Friend*, *Nike+*, *Xively* or *Evrythng*.

E. Artificial Intelligence in social networks

Facebook, *Twitter*, *LinkedIn*, and others are beginning to use artificial intelligence techniques to build their "deep learning" capacities. They are starting to process all the activity occurring over their networks, from conversations, to facial recognition, to gaming activity. Advances in cutting-edge artificial intelligence research, which program machines to perform high-level thought and abstractions, are helping social networks and their advertisers get insights from unstructured consumer data.

And one area demanding further progress is the Human-Computer symbiosis as represented by such experiments as IBM's Watson [15] and IpSoft's Amelia [16], whose stated purpose is extending a human's capabilities by applying intelligent artificial systems techniques, such as deep learning and social and interpersonal communication.

F. Crowdfunding

Crowdfunding, which can be likened to donations, is a capital collection method where common people, and not necessarily professional investors, could fund small personal or business projects by putting their own money into a kind of collective account. Originating from the evolution of Social Web technologies, crowdfunding has gained a followship because of its simplicity, and by removing more formal and traditional forms of loans, like those provided by banks, out of the picture. The most well-known crowdsourcing sites are: *Kickstarter*, *Indiegogo*, *RocketHub* or *Razoo*.

G. Crowdsourcing software development

Voluntary contribution to the creation of new software products, and amelioration of existing versions, is also a recent phenomenon originating from Social Web interaction. Crowdsourcing of software development implies the participation of large numbers of what could be termed a multidisciplinary team involving from designers, to IT architects, to code developers, to relational and documental databases administrators and developers. It is a paradigm shift from industrial mode to peer production mode with a clear impact on both time and money needed for the implementation of an IT product. By having a crowd of volunteers available, testing is much more thorough, and possible issues are detected earlier and then corrected, to the overall benefit of the community of interested parties, and consequently quality is enhanced. This collaborative software development model has now a very widespread use and all of the Free and Open Source Software (FOSS) initiatives pursue the same objectives.

One good example of crowdsourcing of software development are mashups. Loosely defined as mixing and matching content from more than one source to create a single new service displayed in a single graphical interface, the earlier uses for mashups were maps on which geolocating pictures and videos. Mashups have been made possible by the common availability of APIs to make a developer's life a lot easier.

VI. CONCLUSION

Social Interaction Technologies (SIT) have had a transformational effect in many aspects of our lives, since they touch many fields and they have impacted on so many fields in a clear convergence both socially and technologically. Daily activities of many businesses are being socialized, incorporating the central topics of social software (create, connect, contribute, and collaborate) into a multidisciplinary ecosystem of interactive and networked computing.

We have reviewed a number of social interaction tools and some special usages where they have shown a greater effect and impact. The economic results of the so-called social media economy have yet to be produced, but only in terms of productivity increase, and employees and customers satisfaction, the value is certainly remarkable.

REFERENCES

[1] T. O'Reilly (2005). Available: http://www.oreilly.com/pub/a/web2/archive/what-is-web-20.html

[2] J.C.R. Licklider. "Man-Computer Symbiosis" (1960), *IRE Transactions on Human Factors in Electronics* (a copy can be accessed at http://worrydream.com/refs/Licklider%20-%20Man-Computer%20Symbiosis.pdf)

[3] M. Mitchell Waldrop (2001). *The Dream Machine: J. C. R. Licklider and the Revolution That Made Computing Personal*. New York: Viking Penguin

[4] J. Hummel, *Online-Gemeinschaften als Geschäftsmodell: Eine Analyse aus sozio-ökonomischer Perspektive*, 2005, Deutscher Universitäts-Verlag.

[5] Mochón, F., and M. Rojas, "IJIMAI Editor's Note - Vol. 2 Issue 5", International Journal of Interactive Multimedia and Artificial Intelligence, vol. 2, issue Special Issue on AI Techniques to Evaluate Economics and Happines, no. 5, pp. 4-5, 03/2014

[6] Calderón JCP, et al., Informe sobre la situación del Sector TIC 2010 a 2013: Mirada global y de España, Sociedad y Utopía. Revista de Ciencias Sociales 43, 84-101

[7] S. Milgram (1967) "The Small World Problem", *Psychology Today*, Vol. 2, 60–67

[8] MMORPG gamelist: http://www.mmorpg.com/gamelist.cfm

[9] P. Merholz (2004) *Metadata for the masses*. (http://www.adaptivepath.com/ideas/e000361/)

[10] T. Vander Wal (2004). Folksonomies (http://vanderwal.net/folksonomy.html)

[11] A. Toffler (1980). *The Third Wave*. New York: Bantam Books.

[12] R. Putnam (2000). *Bowling Alone: The Collapse and Revival of American Community*. New York: Simon & Schuster
[13] F. Fukuyama (1995) *Trust: The Social Virtues and the Creation of Prosperity*. Cambridge University Press.
[14] W. Richardson. (2006). *Blogs, wikis, podcasts, and other powerful web tools for classrooms*. Thousand Oaks, CA: Corwin Press.
[15] IBM?s Watson: http://www.ibm.com/smarterplanet/us/en/ibmwatson/
[16] IpSoft's Amelia: http://www.ipsoft.com/what-we-do/amelia/

Teófilo Redondo (BArts -1985, MArts - 1986; Universidad Complutense de Madrid - UCM) is Project Portfolio Coordinator at Zed Worldwide, in the Deparment of Innovation. He was before Technology Architect & Director of Innovation Projects at Universidad Internacional de La Rioja (UNIR). Previously he developed a career at IBM covering several areas like Cloud Computing and Big Data Architectures, Enterprise Solutions Architect (SAP, Oracle Solutions, Dassault Systemes), and as SOA Architect. He started in the Research Division of IBM with several projects on Machine Translation, during which he produced a number of articles on this subject. He was Visiting Scholar at Stanford University (1987). He currently holds an Associate Professorship at UNIR teaching about eLearning in Social Networks as well as Natural Language Processing techniques. He is affiliated with SEPLN (Sociedad Española para el Procesamiento del Lenguaje Natural) since almost the beginning.

New Challenges on Crossplatform Digital Contents

Jesús Iglesias Feijoo, Guillermo Amat Gomariz

Zed World Wide

***Abstract* — When we speak about devices and platforms, generally we think about those of general use which are currently available (mainly smartphones and tablets). Surely, we would forget all those which are on the way (watches, glasses, cars) and those which are coming. The Internet of Things will transform the technological world in which we are into an amalgamation of devices and interfaces. This paper analyses the challenge for the coming years of getting all these new devices to communicate between them, regardless of their technology and the platforms they use, and it is based on the works done under the Visio Project, funded by the Spanish Ministry of Industry, Energy and Tourism. Finally, a truly universal platform to avoid market fragmentation and provide access to information and services is proposed.**

***Keywords* — Apps, cloud, crossplatform, internet of things**

I. Introduction

OVER the past 20 years we have been concerned about the compatibility of classic multimedia content (images, music, and video). It has never been as easy as it is now to play any media format on most popular devices (computers, smartphones, tablets...). This fact has been reached due to a slow process of technical improvements, format standardization processes, and also the adoption of these advanced and standardized formats by manufacturers and developers. This has led to a multi-device and multi-platform media scenario. The achievement has been to ensure the user is able to play media on any device, without needing to have the same content in multiple formats to display on different devices, and thus, avoiding the case of having to pay for every required format.

As an example of the opposite, in the late 90s, although HTML was a widely adopted standard, the emergence of various web browsers led to a real browser war where the most affected were web page developers, as they had to create an almost tailored version of their work for each of these browsers. Even worse, the end users themselves were affected as they lost some freedom regarding what browser to use, because some of them were unable to view certain web pages with certain browsers.

One more case to mention is the appearance of the Apple Store in July 2008 [1]. The iPhone application market revolutionized the market of content for mobile devices. It was the beginning of a revolution but also the beginning of the same old mistake. The rise of this concept and the spread of smartphones have achieved something particularly interesting, since nobody but the developers seem to be concerned about the total lack of cross-platform support for these new multimedia contents: applications.

This document is structured as follows: Section II defines the current problem of content incompatibility, section III explains the implications of security issues in the current scenario while section IV introduces the relation with cloud computing. In section V some existent partial solutions and approaches are explained. Finally, section VI proposes a number of possible solutions to achieve a cross platform of digital content environment.

II. Current problem

The next section shows that regardless of the type of Smart Object, the problem is common for all of them. It is observed that the strategy designed for smartphones has propagated the same problem to the new technological sectors, as they have all copied the same paradigm of proprietary, closed and completely isolated systems.

A. Smartphones

As these new "application markets" evolve, the various players in terms of operating systems are making their application platforms available for developers, thus consolidating two major companies (Apple and Google) over the rest of competitors:

1. iOS
2. Android
3. Windows Phone
4. Blackberry
5. Others

From the user's point of view, this behavior has created a barrier against freedom of choice of the user device. This occurs, for example, when users buy applications for their smartphones. If, after a few months, they decide to switch to another device with a competing platform, e.g. iOS to Android, then what happens is that previously purchased

DOI: 10.9781/ijimai.2015.323

applications must be acquired again. When a user has to pay again for all his applications, the decision to change one platform for another is a deterrent like no other, as the cost after several years of use, paying for several applications, can be very high and unaffordable.

As a consequence, currently we do not buy applications but the right to use one in certain platform. A similar fact occurred years ago with the attempt to encode media formats using DRM technologies [2], trying to restrict music and videos to certain players. This, completely absurd from the point of view of the user, is however a big deal for the owners of the application stores.

Therefore, a problem for the development of multiplatform content arises. This is what usually is mentioned to explain the lack of such content. But on the other hand, this deficiency makes it difficult to find useful multiplatform content to end users for their consumption. However, the latter, as explained above, is not a serious problem for big players, as it is yielding great benefits for them.

B. Smart TVs

As for Smart TVs, they are undergoing a similar process. Many manufacturers started to launch their new devices based on Android OS but now major TV brands are introducing their own operative systems, aiming at differentiating them from the rest of their competitors:

- Samsung has recently announced Tizen OS for Smart TVs [3].
- Panasonic is going to be using Firefox OS [4]
- LG is already using Webos [5].

Currently, Sony is the only one which is still faithful to Android. The rest of manufacturers have realized that, by doing that, they are increasing Google's profits.

The main conclusion is that we are witnessing a big fragmentation of operating systems for future TVs. This is not necessarily a negative thing; on the contrary, it increases competition and thus, the potential benefits for users. However, this fragmentation means, again, a lack of multiplatform content to allow a total availability of applications. In other words, each manufacturer will have to worry about increasing their applications catalogue on their own.

Smart TV content developers are in the same situation that occurs for the kind of devices that were previously discussed. Anyone who wants to make an application for Smart TVs will be forced to repeat the development as many times as platforms they wish to reach.

C. Connected Cars

Connected cars are one of the markets that most excitement and growth will experiment in the coming years. Traditionally car interfaces have been completely created by each car manufacturer. There have not been major horizontal suppliers of user interfaces. As the vehicles were implementing dashboard displays and infotainment systems, each manufacturer was developing their own UI to suit their needs. Therefore, there is no relationship between one car manufacturer and the rest. Now it is time for these interfaces to evolve providing connectivity, allowing the use of applications and so forth. The conclusion is that this sector has arrived to the same point we explained before: each manufacturer has their own proprietary and closed system and there is no way to get some multiplatform content.

In recent times, this situation has become increasingly complex due to new products being launched by two of the main mobile communication companies: Apple for Car Play [6] and Google for Android Auto [7]. MirrorLink should also be added to this list, as it is the solution proposed by most automotive-related manufacturers subscribed to the Car Connectivity Consortium, except for Apple. Regarding its functioning, these three systems show certain similarities, as they all project the Smartphone screen onto the infotainment system's and allow users to use certain applications, although the full catalogue of installed applications in such devices would not be accessible. All these applications would be adapted to be used in vehicles in accordance with the design, safety and usability standards of the car manufacturing sector.

Therefore, it must be stressed that these systems are not native to the actual vehicles. In fact, despite the fact that only a few commercial proposals including them have been pitched, the vehicles that have them will also count on the classic native system as devised by the manufacturer and also on a button that will grant the user access to these solutions. However, in principle, unless manufacturers design their own customization layers for these systems, users will not be able to control aspects such as the vehicle's air conditioning system or the radio, or to access the vehicle's setup panel. To sum up, users will always have two different systems in their vehicle, both with different interfaces.

It is widely thought that these systems may be the ultimate and universal solution concerning the use of applications in vehicles, but this might still be an unrealistic idea if we take into account what these products can actually achieve and the international safety standards to which they are subjected. Besides, it would also be necessary to bear in mind what the most common consequences to the use of these proprietary solutions might be. What would happen if a user decided to purchase a new telephone which they would not be able to use in their vehicle? Or if different members of the same family used the same vehicle but all of them had devices operating on different platforms? If, in a best case scenario, it were possible to install several systems in the same vehicle, what would the total cost be? The whole picture is becoming considerably more complex than it might have seemed in the first place.

D. Smart Objects

The Internet of Things and its Smart Objects, small devices that carry out specific functions within a wider network, are the ultimate development in the technological world. In order to understand their usability and future expansion, we can list a number of examples. For instance, there could be Small Objects that are able to measure the temperature or the relative

humidity index in a certain environment, but there could also be other more elaborate ones that can detect a presence or, to a higher degree of sophistication, help control or prevent diseases. All these devices are connected within wider networks which receive the information gathered by the Small Objects and allow us not only to monitor their activity, but also to carry out actions on the devices so that they respond in real time to the measured conditions.

According to a study by Cisco[9], in 2015 there will be twenty billion connected devices, and this figure could be doubled by 2020. Therefore, Smart Objects are one of the subjects of interest within our industry, since they will allow nearly any everyday device to connect to the Internet and to be incorporated into the communication structures in our home or in our business. These objects would be useful both in an industrial or corporate environment and in a domestic one. In fact, DIY (do it yourself) objects will become one of the main factors contributing to the expansion of IoT networks, along with a decrease in the price of electronic components and the standardization of 3D printers.

Even though Smart Objects could still be considered a brand new element in the grand scheme of things, there is a series of mistakes that are already being made. There are not any standards in place regarding communication and interconnection protocols. There are not any free platforms that allow object networks to be developed universally in a controlled and safe environment. The market is currently monopolized by private solutions that do not interact among themselves, without the existence of open protocols or APIs that make it possible to develop a more sophisticated product.

Let us take the example of a simple "smart bulb" to help us illustrate this situation. The supplier would provide us with an application for our Smartphone from which we would be able to set up a timer for the bulb to go on and off. Thus, when we got home our Smartphone would detect the presence of our lightbulb and, depending on the time, the application would establish whether the bulb needs to be on or off. This simple case is perfectly valid, although it would be far more convenient if we could have a bulb that went on or off according to a much more realistic factor such as the degree of luminosity rather than the time of day. In this case, we should somehow communicate with the bulb and with a sensor. Complications could arise, such as the lack of a common communication protocol between the bulb and any other device, like the sensor. To sum up, nowadays it is difficult to develop Smart Objects networks by different manufacturers and which have a common platform that allows communication amongst them all, carrying out data analyses and making decisions.

We have, once again, come to the same conclusion: the new sector is making the same mistakes when they decide to hinder standardization and the use of open protocols.

III. Security and privacy

Another important subject that seems to be overlooked for the abovementioned parties is that of security and privacy. Manufacturers will launch devices with the sole objective of making our lives easier but, at the same time, these devices gather highly sensitive information concerning users and their habits, which could potentially be dangerous if a third party accessed them with malicious purposes.

The exploitation of vulnerabilities also affects new devices. Let us give as an example the registered case of hacked televisions [10], by which a hacker would gather information by means of the built-in webcam that these televisions incorporate. This will soon extend to connected cars, in which everything is controlled by means of sensors and servos, resulting in hackers being able to access the braking or the accelerating system, among others [11]. Recently, a BMW executive warned us about the interest that many companies are showing lately on the enormous amount of data issued by modern vehicles [12].

A good example of such behavior on the part of manufacturers is that of traditional routers connecting households and small and medium-sized businesses to broadband networks. A decrease in their price resulted in less investment on the development of their firmware, and therefore on their security and privacy mechanisms, which resulted in our current scenario, with millions of routers sold in the last few years presenting great vulnerabilities, since sales were more important than security [13].

To sum up, given the insufficient quality of security nowadays, it is a matter of time before Smart Objects get hacked. This situation is due to manufacturers having ignored any engineering processes concerning the design of security for their devices. This result in a reduction of the costs incurred in during the development of the product and enables a higher number of products to be launched in a much shorter period of time.

IV. Cloud computing

The Cloud is an essential component of any application and, according to analysts, the Cloud services market will grow exponentially in the next few years, along with the Internet of Things industry. However, the use of the Cloud has been limited to simply acting as a data storage application, granting said information an "apparent ubiquity", whereas in actual terms it does not provide any advantages for the end user. The Cloud, along with other broadband data lines, smartphones and tablets, has changed the way we create and use content. Along with its flexibility and pay-per-use model, one of the greatest advantages of the Cloud is its actual ubiquitous access to applications and services on the Cloud. However, nowadays this ubiquity is very much limited to its ability to access applications and services adapted to multiple devices, as these solutions can be:

- Expensive, for they require complex and specific developments for each platform

- Incomplete, because, although there is some degree of adaptation of the graphic interface to each device, it is not a dynamic adaptation and, more often than not, a change of device will imply a change in the model of interaction. The majority of services in the Cloud are useful in terms of their server functionality without user interaction, as most of them focus on infrastructure.

If we looked back a decade ago, we would be able to clearly identify the major software developers worldwide. However, nowadays young people's skills at monetizing their innovation and creativity go mostly unnoticed. With technologies like the Cloud, html5 and other capacities, the application economy has changed dramatically. Amateur programmers can now make use of their creativity and innovative spirit to develop new applications or services and offer them through app stores. Unlike in past decades, today it is not necessary to count on a major capital investment in order to start writing code and start selling new applications or services. Costs have decreased drastically as a result of cloud computing and Infrastructure as Service (IasS). The Cloud has changed the way modern-day software is created.

This ubiquity becomes apparent because, in effect, when sharing data among several applications in different devices and from any location, it becomes clear that we have universal access to that information. The context of the knowledge of applications resides on those data in the Cloud and not on the actual application that has been physically installed on a device. But this "universal" availability is a first step that gets taken for granted by any user, it is the bare minimum, but does not go any further than that. The Cloud at a user level is clearly underused, since it does not allow users to go one step beyond, it cannot be used somehow to fill the niche of applications that are independent of platforms or devices. However, we are sure that there would be many more possibilities of use. Later in this article we will present a number of proposals that may contribute to the abovementioned ubiquity not to refer to data only, but also to the actual user interfaces.

V. CROSSPLATFORM DEVELOPMENT

The current model for application development is not resalable, for a developer will have to create as many versions of their application as platforms exist in the market. This not only means smartphones, as we have seen, but any other type of electronic device that is connected.

The need to develop different versions of applications and to adapt them to different devices entails a high cost. Adaptability is essential in order to create a sustainable application ecosystem.

Currently, the different operative systems are at war, which will result in the need to create more versions of the same application if the ultimate goal is to make said applications available to each and every user. This, however, would not be viable in the current scenario. Fragmentation thus becomes unsustainable.

In such a scenario, developers end up having to sacrifice a number of versions for the sake of the two or three versions of their application that would reach the highest number of potential users, thus maximizing monetization. The remaining options become automatically discarded.

The user interface management system that traditional operative systems use (Windows, Mac, Android, etc.) gives programmers more freedom through the use of its APIs, so much so that it enables them to communicate with the user by setting up interactive objects such as buttons or menus with windows and dialogues that are transferred to the user in order to show them the new information. Once the user has filled in the information required by the application, the window or user process dialogue moves on to validate the received information.

This paradigm for the development of user interfaces is not sustainable in a complete cross-platform environment, since the interaction device followed by the user at any given time does not guarantee the existence of windows or buttons (let us imagine, for instance, that the device were only able to recognize and utter speech).

We cannot state, however, that there are not any solutions in the market that advocate for a cross-platform development environment which would allow the coveted approach of "develop once, deploy many times". These, though, are still limited to specific platforms and services. For instance, there are a number of development environments that enable us to automatically export an application into iOS, Android or Windows Phone, to name a few. There are some systems that may include even more versions. However, there are not any universal solutions that are not focused on smartphones only and that apply to all kinds of devices, as we have mentioned: TVs, cars or any kinds of Smart Objects.

If we look beyond the actual devices, we can find even more problems when it comes to dealing with different interaction systems. An application that has been designed to be used in smartphones will not be easily adapted to a TV and the user experience would not be suitable, either. Moreover, if we wanted to use the same application in a small Smart Object, we would be faced with more frustration and problems related to subjects ranging from the actual human-machine interaction (haptic, voice, gestural) to the visualization of results of said interaction. It cannot be expected to use an application with a user interface if we do not have a screen to see it on.

VI. SOLUTIONS

If we take into account both conceptual and lexical design when proposing universal access to the knowledge base of an application, we can state that, even when accurately planned, said universality cannot be guaranteed, as there will be situations where the interface has not been designed for its general use. This is why it will be necessary, in most situations, to develop several versions of said interfaces which can be adapted to the needs of both devices and human groups.

In order to avoid this, said conceptual and lexical levels have to be dynamically generated according to the contextual and interaction requirements of that moment, creating a user interface management system that is intelligent and able to design and construct interactive dialogues in real time that can be especially adapted to the cognitive, perceptive and motor characteristics of an active user, as well as to the technical characteristics of the device that the user is using at that very moment, granting a great variety of users and devices full access to the interface, including those users who suffer from certain disabilities. Thus, the functionality of the application would be completely independent from the interface, which would hugely increase the system's efficiency in terms of usability and user experience. The computing capacity of the Cloud would be a solution to potential load issues in small devices, whose only concern would be the user interface.

As it is impossible to know about the device's interaction characteristics when designing and compiling, a change of paradigm becomes necessary, where the programmer would determine their needs in terms of UI (types of data required to launch a process, in/out parameters, response, etc.) and an independent service would decide in real time how to communicate with the user and their interaction device in order to obtain the information required. We can now introduce a new concept: the concept of Cloud UI, user interfaces in the Cloud.

Nowadays, we are able to maximize the Cloud's calculus power and the ubiquity that it provides us with in order to use them for the greater benefit of users and applications alike. The goal should be that applications become an essential actor in the user interface, rather than just a final solution regarding data treatment and visualization. This idea is not new, however, and there exist solutions and R+D projects based on these ideas.

Active Video[14], with its platform CloudTV, proposes a solution of this kind for TVs and set top boxes, and it is based on transforming the user interface into a video stream. This would enable the user to visualize the user interface from their device and interact with it as if it were a native application. The difference lies in the fact that the data issued throughout the interaction are sent to the Cloud, processed and sent back to the user as a response in the shape of a new video stream.

With the new interface model in the Cloud, we achieve the coveted paradigm of cross-platform applications. Regardless of the device's operative system, brand or model, we only need a small connector that would enable us to automatically use all the applications that are designed to work on such a platform. It would not be necessary to replicate and adapt applications to different systems; it would only have to be done once.

Another element to discuss would be the fact that user interfaces, instead of being tailored to needs, should become description systems. The VISIO project, developed jointly by the University of Oviedo and Zed Worldwide, is based on this concept. Communication with the user is based on a set of minimum requirements in terms of the application's functionality, which are as follows:

- The information the user must have
- How they must have it
- How it should be communicated to the system

Bearing these in mind, the actual user interface gets relegated to the background and what becomes important is the fact that any person would be able to use a specific application regardless of the device or the kind of visualization and interaction interface they are using. Through the use of mechanisms that describe an application's interface and its behavior towards the user, it would be plausible for the same application to be used in completely different devices:

- Smartphones, Smart TVs, connected vehicles and even small Smart Objects,
- Devices with a screen with, for example, small LED displays or even devices without a screen where an interface reads and describes the application or emits sound effects when the task has been processed successfully.
- Devices with various interaction systems: haptic, gestural, voice, etc.

To automatically generate the applications' interfaces in real time would enable us not only to make a general use of them, as we have just seen, but also it would improve the quality of the user experience by gradually adapting the adaptations and/or alterations of the interface to the user context so that they do not occur suddenly. For instance, through the use of logic engines, a rule defines the size of a button as BIG if the degree of driving precision is LOW and the degree of visual precision is HIGH. The engine defines the probability of an interactive object to be LONG as a combination based on the probabilities of the engine precision and the visual precision. Interaction with touchscreens in work or home environments will be completely different to the interaction that takes place in a moving vehicle (such as a car, metro or bus). These situations could be detected (based on the sensors of the actual device) and we would be able to apply various methods of interaction to each of them. Regarding interaction in movement, the degree of precision for the user suffers significant degradation and this would be detected by the system. When we apply logic rules, when we increase the probability of the driving precision to be low, the chances of generating a bigger size button increase. This is exactly what would be expected in the context we have described, since the use of longer buttons increases the level of precision of the interaction in a moving environment. Once the user has left said environment and the driving precision increases once again, the size of the buttons decreases again.

This simple example that appears to be obvious is not viable to be used with the current user interface design paradigms. All the elements are static and would appear in the same format as they have been previously defined, regardless of the conditions of the user context. Again, there is a seeming ubiquity, but it is completely de-contextualized.

User interfaces tend to be designed to satisfy the needs of the "typical user". The classic design of interfaces does not

consider human individuality and diversity. As a consequence, the resulting systems present serious flaws regarding user experience, since those who do not fall into the category of "typical user" will face problems and frustration on using these systems. The use of systems based on the definition of interfaces would be a solution for this problem: the adaptation for users with disabilities. Applications would be possible to adapt to visually impaired users or to users suffering from reduced mobility. The actual system can be adapted, for it has not been designed to be used in a specific way, but thinking of its utilization, of the information that it must know and of what the data exchange with said system should be like.

However, we must not claim that this would be a perfect, problem-free solution. There is a major problem that is quite obvious: the total dependence of the system on connectivity, without which it would be impossible to render the user interface. Despite the advantages that the use of local caches might have, as well as the possibility to use the offline mode, it becomes quite clear that this system might still be proposing a limited scenario. In any case, we must also take into consideration that the vast majority of applications we use today, despite unfolding interfaces locally, depend almost entirely on a connection to be able to operate. The apparent ubiquity we mentioned earlier on this paper and that stems from the synchronization of data from the Cloud would still be hindering the optimal use of the application, which, even when it is clearly a weakness, does not necessarily mean that native applications are entirely free from it.

Along the same lines, another issue could be a delay in the network that could cause great frustration to the user if there is not an almost immediate response between call and return. Once again, this is a situation that might well happen today, as our dependence on Cloud services is almost total.

As per the advantages of this system, besides the ones that we mentioned in a previous section of this paper, i.e. the fact that they can be cross-platform, used in multiple devices and for general purposes for every type of user, regardless of their condition, we must mention other additional benefits:

- Scalability. It is scalable not only at a computational level, but also at a functional level. Given the fact that all the processes reside in the Cloud, it is relatively simple to add new models of interaction. This scalability applies equally to the number of devices that can be connected to our system.
- A much more efficient user experience. Improvements in the system can be universally applied to all users and devices. With a numerous user contingent we can undertake mass usability tests, significantly increasing the processes of interaction and thus generating functional improvements.

VII. Conclusion

Every strategy designed to reduce the human and technological limitations and achieve a significant increase in the potential use of any technological tool should include the creation of a wide range of communication channels through which humans and machines can exchange information. According to the type of user or device that makes use of the service, we can adapt not only the communication channel to be used but also the way in which information is perceived by the user, i.e. the interface, using even the knowledge that the system might possess about the user context.

The human-machine interaction must be considered as something more than the simple use of a touchscreen and comprehensively and universally include any type of device, interaction and human condition, so that technology becomes a means to unite, not to separate, for those who use different platforms, for those who suffer from certain disabilities or for those who, in certain situations, cannot use certain types of interaction.

Even though it is down to users to demand these technological advances, it is also the responsibility of device and operative systems manufacturers to approach technology with a long-term vision, through the use of open standards and protocols, thus facilitating benefits for all the parties involved. It might be that today's business models, based on pay-per-download, or even in-app sales and subscriptions that are inaccessible for providers in app stores are not ideal and should be looked into. It may also be convenient to test new models that, in accordance with the ubiquity requirements that future electronic devices might present, continue to allow providers to have satisfactory account balances.

An understanding approach to the user context in the upcoming years will become one of the major axis of online sales channels (particularly from mobile devices), offering the user what can be inferred that they are going to need and/or want according to who they are, where they are and their previous activity with information systems. If these means can be used to "recommend" the purchase of specific goods and services, why should they not be used to improve the interface user experience through the use of contextual ubiquity?

Total ubiquity is advantageous not only for users themselves, but also for any sphere of their everyday lives. It is not a technological matter, but a social and ethical one. The Internet's neutrality is a widely known concept and this neutrality should be made extensive to applications and services, since these represent the digital profile of every individual. The owner should thus be free to export said profile to any platform that they deem appropriate.

ACKNOWLEDGEMENTS

This paper is being supported by the VISIO Project (http://www.bitmonlab.com/rdi/Visio/index.html), funded by the "Acción Estratégica de Economía y Sociedad Digital" programme of the Ministry of Industry, Energy and Tourism of Spain. The VISIO project is collaboration between the University of Oviedo, Spain and Zed Worldwide, Spain.

References

[1] Apple App Store press release, July 14, 2008.

[2] http://www.apple.com/pr/library/2008/07/14iPhone-App-Store-Downloads-Top-10-Million-in-First-Weekend.html
[3] Removal of restrictions can decrease music piracy
[4] http://business.rice.edu/jgsb2011_2column.aspx?id=4294967941
[5] SAMSUNG Revolutionizes the Viewing Experience With Innovative New SUHD TV, January 05, 2015. http://www.samsung.com/us/news/24393
[6] Panasonic and Mozilla to Collaborate on New Firefox OS Open Platform for Next Generation Smart TVs, January 6, 2014. http://www2.panasonic.com/webapp/wcs/stores/servlet/prModelDetail?storeId=11301&itemId=705511
[7] LG Electronics Acquires webOS from HP to Enhance Smart TV, February 2, 2013. http://www.lg.com/us/press-release/webos-release
[8] Apple Car Play. https://www.apple.com/ios/carplay/
[9] Android Auto. http://www.android.com/auto/
[10] Mirrorlink. http://www.mirrorlink.com/
[11] The Internet of Things. How the Next Evolution of the Internet Is Changing Everything. Dave Evans, Cisco, Cisco Internet Business Solutions Group (IBSG), April 2011. http://www.cisco.com/web/about/ac79/docs/innov/IoT_IBSG_0411FINAL.pdf
[12] Samsung Smart TV a spy in the living room as webcam hack revealed, August 2, 2013. http://www.slashgear.com/samsung-smart-tv-a-spy-in-the-living-room-as-webcam-hack-revealed-02292655/
[13] Hacker Says Attacks On 'Insecure' Progressive Insurance Dongle In 2 Million US Cars Could Spawn Road Carnage. Forbes, January 1, 2015. http://www.forbes.com/sites/thomasbrewster/2015/01/15/researcher-says-progressive-insurance-dongle-totally-insecure/
[14] BMW sounds alarm over tech companies seeking car data. Fiancial Times. http://www.ft.com/intl/cms/s/0/685fe610-9ba6-11e4-950f-00144feabdc0.html#axzz3PG1kItGq
[15] Lizard Stresser Runs on Hacked Home Routers. Krebs on Security, January 2015. http://krebsonsecurity.com/2015/01/lizard-stresser-runs-on-hacked-home-routers/
[16] Active Video. http://www.activevideo.com/

Jesus Iglesias Feijoo has worked in Software Development over fifteen years. After creating his own Internet company in 2000, he joined Zed Worldwide where he has been leading international software projects during the last ten years. Nowadays he works as r&d project manager and leads an innovative development team. He maintains his entrepreneur activity developing successful apps for several platforms.

Guillermo Amat Gomariz was graduated in Computing Science in the Universitat Jaume I, Castellón, Spain, in 2003. He is a R&D Project Manager at Zed World Wide where he has lead several Smart City projects. His major research interests are Geographical Information Systems applied to Smart Cities, the Internet of Things and software portabilityInformation Systems applied to Smart Cities, the Internet of Things and software portability.

Empirical Comparison of Graph-based Recommendation Engines for an Apps Ecosystem

Luis F. Chiroque[1], Héctor Cordobés [1], Antonio Fernández Anta[1], Rafael A. García Leiva[1], Philippe Morere[1], Lorenzo Ornella[3], Fernando Pérez[2], Agustín Santos[1]

[1]*IMDEA Networks Institute, Madrid, Spain*
[2]*U-Tad, Madrid, Spain*
[3]*Politecnico di Torino, Turin, Italy*

***Abstract* — Recommendation engines (RE) are becoming highly popular, e.g., in the area of e-commerce. A RE offers new items (products or content) to users based on their profile and historical data. The most popular algorithms used in RE are based on collaborative filtering. This technique makes recommendations based on the past behavior of other users and the similarity between users and items. In this paper we have evaluated the performance of several RE based on the properties of the networks formed by users and items. The RE use in a novel way graph theoretic concepts like edges weights or network flow. The evaluation has been conducted in a real environment (ecosystem) for recommending apps to smartphone users. The analysis of the results allows concluding that the effectiveness of a RE can be improved if the age of the data, and if a global view of the data is considered. It also shows that graph-based RE are effective, but more experiments are required for a more accurate characterization of their properties.**

***Keywords* — Recommendation engines, smartphone apps, graph theory, collaborative filtering, flow algorithms.**

I. INTRODUCTION

A. Motivation

It is becoming very common in online platforms (shopping websites, online newspapers, online social networks, smartphone apps, etc.) to recommend items to the users that will (hopefully) be of their interest. This trend is becoming so general that Anderson predicted that we are "leaving the age of information and entering the age of recommendation" [4]. The items to recommend are selected by a recommendation engine (RE) that typically leverages the user profile, the context, and historical data. The RE typically has a catalog of items from which to choose its recommendation, and there are spaces in the online platform viewing area in which the recommended product is presented. The context of the user typically includes its past navigation history, including the current viewing context, which may involve a product (e.g., in a shopping website), a piece of news (e.g., in an online newspaper), a user profile (e.g., in an online social network), or the application that is being executed (e.g., in a smartphone).

Recently, the most popular algorithms used in RE are based on collaborative filtering [15]. This technique makes recommendations based on the historical data of all the users and the estimated similarity between them. Typical metrics used for the computation of customers' similarity include Pearson correlation coefficient, adjusted cosine similarity, Spearman's rank correlation coefficient, and mean squared difference.

In parallel with the advances in RE algorithms, we have observed that graph theory and network analysis has been useful in different contexts to extract information from data. This information is not an explicit part of the data, but it is implicitly contained in the underline structure. Examples of this approach are the use of pagerank to identify the most relevant web pages [8], or a recent use we have made of graphs to classify tweets [9]. We believe that graph theory and network concepts can also be useful in the context of recommendation.

B. Contributions

In this paper we present an exploratory work on using graphs to build RE. We have devised, developed, and evaluated RE based on collaborative filtering to promote an ecosystem of smartphone apps. In this ecosystem, the users of the apps get banners advertising other apps that they have not installed (yet). The objective of the RE is maximizing the click-through rate (CTR) of users in these banners, and maximizing the installation of new apps. In addition we have devised one particular RE to promote a specific subset of apps. The proposed RE create models of the ecosystem as networks formed by apps, and use graph theoretic concepts like edges weights or network flow.

The performance of the RE proposed has been evaluated in a real apps ecosystem. Several years' worth of historical data has been used to create the networks that model the ecosystem.

DOI: 10.9781/ijimai.2015.324

Then, using them, the different RE were put to work with real users for about a week. The analysis of the results obtained has shown big (statistically significant) differences in CTR and installation success of the different RE. The results allow concluding that the effectiveness of a RE can be improved if the age of the data, and if a global view of the data is considered. It also shows that graph-based RE are effective. However, some of the results are puzzling, and hence more experiments are required for a more accurate characterization of the properties of the proposed RE.

C. Structure

The rest of the paper is structured as follows. In Section II we present the problem to be solved. In Section III we describe the RE we have proposed and that will be evaluated in this paper, with the underlying graph they use. In Section IV we present the experiment we have conducted, the results obtained, and some discussion on them. Section V presents previous work related to this paper. Finally, Section VI concludes the paper.

II. Problem Statement

As described, we have a *smartphone app ecosystem.* In this ecosystem, a user that is running an app (called the *publisher*), gets *banners* advertising other apps of the ecosystem it has not installed yet. The objective is to devise a RE that tells the system which app to advertise to a given user at a given time, possibly as a function of the user and the publisher, in order to achieve one (or more) of the following objectives.

- *CTR Maximization:* The objective is to maximize the number of times the user clicks in the banner to get more information about the apps advertised.
- *Installations Maximization:* The objective is to maximize the number of times the user installs the app advertised.
- *Targeted Promotion:* The objective is to maximize the number of times users install a preselected set of apps to be promoted.

An initial hypothesis we will make is that, once a user has clicked in a banner, the probability of installing the corresponding app is roughly same. This has made us concentrate initially in RE for the CTR Maximization and Targeted Promotion objectives. (As will be seen from the results obtained, this initial hypothesis needs some revision.)

III. Recommendation Engines Proposed

In this section we describe the recommendation engines we have proposed and evaluated in this paper. In order to describe them, we build graphs from historical user data that convey the essential information that is required by the corresponding RE. Hence, we start describing the graphs we need and use, and then we give the algorithms used by the RE to select an app to advertise.

A. Apps Graphs

All the graphs used in this work will have the set of apps A of the ecosystem as vertices. Moreover, all of them are weighted graphs, and the main difference among them is the weights that are allocated to edges. The graphs used are the following.

Shared Users (SU) Graph. The SU graph is an undirected weighted graph $G_{SU}=(A,E,w)$, where $E=\{\{i,j\}: i,j\in A\}$ and the weight $w(e)$ of an edge $e=\{i,j\}\in E$ is the number of users of the ecosystem that have currently both apps i and j installed.

Aged Shared Users (ASU) Graph. The ASU graph is an undirected weighted graph $G_{ASU}=(A,E,w)$, where $E=\{\{i,j\}: i,j\in A\}$ like in G_{SU}. The difference in this case is that the contribution to the weight $w(e)$ of an edge $e=\{i,j\}\in E$ of a user (that has currently both apps i and i installed) is a function of the time the user has had the apps installed. In particular, let U be the set of all users and $U(a)\subseteq U$ be the set of users that have app $a\in A$ installed. Also, let $age(u,a)$ be the time since user $u\in U(a)$ installed app $a\in A$ (in some suitable units). Then,

$$w(\{i,j\}) = \sum_{u\in U(i)\cap U(j)} \delta^{\min\{age(u,i),age(u,j)\}},$$

where $\delta\leq 1$ is the decay factor. (The intuition is that users that installed an app long time ago are less important for the app.)

CTR Graph. The CTR graph is a *directed* weighted graph $G_{CTR}=(A,E,w)$, where $E = A\times A$ and the weight $w(e)$ of an edge $e=(i,j)\in E$ is the CTR observed when banners with app j are presented to the users with app i as publisher.

B. Recommendation Algorithms

Using the above graphs we can describe now the RE considered in this work.

Shared Users.

Let us consider the SU graph described above. Assuming the publisher app is i, the app recommended j is the one that has the edge with i of largest weight. I.e., $j=argmax_k(w(\{i,k\}):k\in A)$. In this case, this means that j is the app with the largest number of common users with i.

The approach of this algorithm is not new, and it is among the first ideas one may think of when resigning recommendation algorithms.

Collaborative Filtering.

This algorithm also uses the SU graph. Given the user to which the banner will be presented, and the set I of applications the user has already installed, the app j recommended is the one that has a largest aggregate weight with those in set I. I.e., $j=argmax_k(\Sigma_{i\in I} w(\{i,k\}):k\in A)$.

Again, this approach is not very novel, since it is common to many collaborative filtering algorithms to use some linear algebra approach that can achieve similar results as this one. For instance, considering the weights of the SU graph as a matrix W, and the applications already installed by the used as a vector v, the algorithm proposed would recommend the app that corresponds to the largest element of the vector v^TW.

Aged Shared Users.

This algorithm applies the same process as Shared Users, but in the ASU graph. As far as we know this algorithm is new.

Aged Collaborative Filtering.

This algorithm applies the same process as Collaborative

Filtering, but in the ASU graph. As far as we know this algorithm is also new.

Maxflow.

This algorithm uses the CTR graph with the objective of promoting a preselected subset *P* of apps. The algorithm takes the publisher app *i* and solves a flow maximization problem [11] from *i* to each of the apps in *P*, where the weight of each link is considered its capacity. Then, it recommends the neighbor of *i* whose aggregated flow is the largest. I.e., imagine that the solution of the maximum flow problem from *i* to $a \in P$ sends $f(a,k)$ units of flow across link (i,k). Then, the recommended app is $j=argmax_k(\Sigma_{a \in P} f(a,k):k \in A)$.

To our knowledge, the Maxflow RE is also new. The intuition behind it is that instead of directly promoting the apps in *P* it is better to promote those that will drive the user to them.

In addition to the 5 RE described, we will consider for reference two trivial algorithms.

Random.

This algorithm recommends an app at random using a uniform distribution over the set of available applications. As we just said, the goal of the random RE is to have a reference with which all the other RE can be compared.

Static Promotion.

This algorithm always recommends one of the applications of the set *P* to be promoted (chosen uniformly at random). It does not depend on the user installed applications, nor the publisher.

IV. Performance Evaluation

In this section we describe the experiment we have conducted in order to evaluate and compare the RE proposed. Then, the results obtained in the experiment are presented and briefly analyzed.

A. Implementation of the Experiment

As mentioned previously, the evaluation of the RE has been done in a real apps ecosystem. This ecosystem is formed by roughly 300 apps with more than 4 million users.

To build the graphs used by the RE and described in the previous section, we have used more than 3 years worth of data. This adds to more than 100 GiB of historical data structure in more than 1.4 billion records. This data has been processed with Big Data technologies (Hadoop, Pig, Hbase [5-7]) in the Amazon Elastic Map Reduce [2] environment. The processing involved cleaning the historic data generated a clean dataset of more than 700 million records of events, containing the user, the publisher, the app advertised (in the banner), the action associated to the event (add, click), and the timestamp.

From the clean dataset just obtained, the above-described graphs were built. The construction of the aged graph ASU used a value of $\delta=0.95$ and the age is measured in units of weeks. It is important to note that the historical data to which we had access did not record explicitly the installation of the apps. The fact that a user had an app installed was extracted from the data because the app appeared as publisher in some event.

Once the graphs were ready, we run an experiment in the real system for about a week (from Jun 2nd, 2014 to Jun 10th, 2014). In this experiment, the different RE recommended the apps shown in banners to the users. In order to avoid cross interference, the same RE generated all the banners for the same user. For the targeted promotions RE proposed (Maxflow and Static Promotion) a manually selected set of 5 apps were chosen to be promoted. At the end of the experiment, banner were shown to more than 300,000 users, and each RE had done more than 130,000 recommendations.

After the experiment, the data of number of banners recommended by each RE, the number of clicks by the user, and the number of apps installed was obtained. It is important to note that the data obtained was cleaned. For instance, multiple clicks associated to the same banner where counted only once. Regarding installations, we assumed that a banner had caused the installation of an app if the app was used (by the user) within 72 hours after the banner was shown.

B. Experiment Results

The basic results obtained in the experiment are presented in Table I. For each RE the table shows the total number of banners that used the RE for recommendation, the number of banners on which the user clicked, and the click through rate, CTR, which is the ratio of the former two values. Additionally, the number of installations from the banner is also shown. Finally, we present two metrics, installation to banners rate (IBR) and installation to clicks rate (IBR), which are the ratio of the number of installations versus the number of banners and the number of clicks, respectively.

TABLE I
RESULTS OF THE EXPERIMENT

RE	Banners	Clicks	CTR (%)	Installs	IBR (%)	ICR (%)
Random	140894	1993	1.41	126	0.09	6.32
Shared Users	133818	2095	1.57	299	0.22	14.27
Aged Shared Users	139417	2258	1.62	390	0.28	17.27
Collaborative Filtering	134790	1966	1.46	329	0.24	16.73
Aged Collaborative Filtering	133623	2204	1.65	375	0.28	17.01
Static Promotion	138922	1929	1.39	215	0.15	11.15
Maxflow	140858	2302	1.63	290	0.21	12.60

As can be seen from the results presented, the CTR observed is different for different RE. Table II presents a comparison of the differences of the CTR achieved by the RE. In each entry of the table it is show (in percentage) the increase in CTR achieved if using the RE of the column instead of the RE of the row. When this number is negative the CTR in fact decreases, and the value is marked in red.

TABLE II
INCREASE OF THE CTR (IN PERCENTAGE) WHEN USING ONE RE (COLUMN) VERSUS ANOTHER (ROW)

	Random	Shared Users	Aged Shared Users	Collab. Filtering	Aged Collab. Filtering	Static Promotion	MaxFlow
Random		10.68	14.50	3.11	16.60	-1.84	15.53
Shared Users	-9.65		3.45	-6.83	5.36	-11.31	4.39
Aged Shared Users	-12.66	-3.34		-9.94	1.84	-14.27	0.91
Collab. Filtering	-3.02	7.34	11.04		13.08	-4.80	12.05
Aged Collab. Filtering	-14.24	-5.08	-1.81	-11.57		-15.82	-0.92
Static Promotion	1.87	12.75	16.64	5.04	18.79		17.70
MaxFlow	-13.45	-4.20	-0.90	-10.75	0.93	-15.04	

Given these differences CTR between the RE used, we want to determine if they are statistically significant. For that, we have computed a *z*-test [19] to compare the CTR of each pair of RE. For a given pair of RE the null hypothesis is that both populations are extracted from equal distributions (and hence the differences are simply due to statistical noise). The alternative hypothesis is then that the distributions are different. We compute the *z* value for two RE (numbered 1 and 2) as

$$z = \frac{p_1 - p_2}{\sqrt{p(1-p)}\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}},$$

where n_i is the number of banners of RE i, $p_i=x_i/n_i$ is the ratio between the number of clicks x_i and the number of banners, and p is the pool population, defined as

$$p = \frac{x_1 + x_2}{n_1 + n_2}.$$

The results of all the z-tests for the CTR are shown in Table III. Using an alpha value of 0.05, we reject the null hypothesis if the z value is outside the interval [-1.96,1.96], with a confidence of 0.95. If the null hypothesis is rejected for a pair of RE, it means that with a probability of at least 0.95 the CTR of one is larger than the CTR of the other. In Table III the pairs of RE for which the null hypothesis is rejected have white background, and we use yellow background for the cases in which the null hypothesis cannot be rejected.

As done with the CTR, we have also computed the *z*-values for the IBR for each pair of RE, with the objective of identifying when the difference in IBR shown in Table I are statistically significant. Again, the null hypothesis is that the differences are only due to randomness. Table IV presents the results, where the white background again means that we reject the null hypothesis with a confidence of 0.95.

TABLE III
Z-VALUES COMPUTED FOR THE CTR OF EACH PAIR OF RE.

	Random	Shared Users	Aged Shared Users	Collab. Filtering	Aged Collab. Filtering	Static Promot.
Shared Users	3.26					
Aged Shared Users	4.44	1.13				
Collab. Filtering	0.97	-2.27	-3.42			
Aged Collab. Filtering	5.01	1.72	0.61	4.00		
Static Promotion	-0.58	-3.83	-5.01	-1.55	-5.57	
MaxFlow	4.76	1.43	0.31	3.73	-0.31	5.32

TABLE IV
Z-VALUES COMPUTED FOR THE IBT OF EACH PAIR OF RE

	Random	Shared Users	Aged Shared Users	Collab. Filtering	Aged Collab. Filtering	Static Promotion
Shared Users	8.93					
Aged Shared Users	11.75	2.93				
Collab. Filtering	10.00	1.11	-1.83			
Aged Collab. Filtering	11.73	2.95	0.04	1.85		
Static Promotion	4.95	-4.13	-7.08	-5.25	-7.07	
MaxFlow	8.05	-0.99	-3.97	-2.12	-3.98	3.18

Finally, we present the results of the different RE as promoters of specific apps. As described above, we have devised a RE, Maxflow, specifically for targeted promotion of apps, and used it for promoting the 5 chosen apps.

In addition, as mentioned, we implemented a trivial RE, Static Promotion, which only recommends the 5 apps to be promoted. In Table V we present the number of installations that these RE achieved for each of the four apps to be promoted, numbered from 1 to 5. For comparison, we also show the numbers of installations achieved with the other RE.

C. Discussion

Table I shows significant differences between the RE used. The first fact to note is that, as expected, both Random and Static Promotion have very low CTR and IBR. All the other algorithms have a CTR that is at least 3% higher than Random and 5% higher than Static Promotion (see Table II). The difference in IBR is even higher, were every RE achieves at least a 133% increase over Random and 40% over Static Promotion.

TABLE V
INSTALLATIONS OF THE 5 APPS PROMOTED

RE	App 1	App 2	App 3	App 4	App 5
Random	2	0	4	0	4
Shared Users	18	1	13	45	31
Aged Shared Users	13	0	15	50	42
Collaborative Filtering	7	0	22	36	39
Aged Collaborative Filtering	13	0	22	47	43
Static Promotion	35	17	38	70	47
Maxflow	0	0	2	11	46

Comparing the CTR of the rest of RE, there is a significant difference between Shared Users versus Aged Shared Users, and Collaborative Filtering versus Aged Collaborative Filtering. This difference leads to conjecture that the preferences of the users change over time. This is the reason why the RE that take that into account this evolution behave well. Somewhat surprising is the high CTR achieved by Maxflow, which has the second largest CTR, since the objective of this RE is not to maximize the CTR.

Table III shows that the differences between the CTR are statistically significant many cases. In particular, in terms of CTR, the *z*-test divides the RE into two groups. One group has CTR that is statistically smaller than the other. Random, Static Promotion, and Collaborative Filtering form the group of low CTR. The group of high CTR includes Shared Users, Aged Shared Users, Aged Collaborative Filtering, and Maxflow. Observe that a larger alpha value in the *z*-test would differentiate the RE further.

Looking at the ICR columns in Table I, we can see that the values in the column differ significantly. This disproves our initial hypothesis that, once a user clicks in a banner, she has a similar probability of installing the app. The conclusion is that it is not enough to aim at maximizing CTR if the objective is to get app installations. For instance this causes that Maxflow is the RE with the lowest IBR from those not for reference. Moreover, the values in Table IV show that this difference is statistically significant. From this table we can conclude that all RE are more efficient in terms of installations than Random. Also, that all the "smart" RE are more efficient than Static Promotion (of course, this is natural since the target of this RE is not maximizing installations).

Table IV also shows that the effectiveness in installations of Aged Shared Users and Aged Collaborative Filtering is higher than the other RE. This reinforces the conjecture that the preferences of users change over time and this changes has to be taken into account by the recommendation system.

Finally, regarding targeted promotion, in Table V we can observe that Maxflow achieves a low number of installations for the 5 promoted apps, especially compared with Static Promotion (but even versus all the other RE except Random). This result is disappointing, and it requires further study. Our conjecture is that the experiment conducted was too short to observe the effect of Maxflow, which promotes the apps that lead to other apps. Other lines to explore are the modification of Maxflow in two ways. First, Maxflow must be tested using an IBR graph instead of the CTR graph (since, from a previous discussion CTR is not the critical metric if we want installations). Second, the graph used by Maxflow must consider aging, since as we have observed this is an important aspect of the data. In any case, another conclusion we obtain from Table IV is that using Static Promotion for targeted promotion of apps seems like a valid option.

V. RELATED WORK

The most common approaches to the recommendation problem can be grouped into three types.

- Collaborative filtering [15]: In this approach users are represented by an *N*-dimensional vector of items, and the recommender looks for users who have similar rating patterns as the target user. Then, it uses the ratings from those like-minded users to make a recommendation for the target user.
- Cluster models: This approach divides the customer base into many segments, and treats the recommendation task as a classification problem. Segments are created using a clustering, or some other unsupervised learning algorithm.
- Search-based methods: In this approach, given the target user's purchased and rated items, the algorithm constructs a search query to find other popular items by the same author, artist, or director, or with similar keywords or topics.

As an example, Amazon uses its own recommendation algorithm, called item-to-item collaborative filtering [17], to personalize the online store for each customer. The algorithm is focused in finding similar items, not similar customers, and hence it scales independently of the number of customers. However, the challenge is to make it scalable with the number of items in the product catalog.

Most of the collaborative filtering algorithms we have found in the literature assume that user preferences remain stable and consistent over time [14]. We believe this is not generally the case, and our conjecture is supported by the fact that in our experiment the RE that considered aging performed very well.

Methodologies for the evaluation of RE have been proposed in [18] and [12]. Other aspects, like advertising effectiveness and Return of Investment (ROI) on social networks, have been a big topic of discussion for advertisers in the past decade [3]. ROI has been typically measured through econometric models that measure the impact of varying levels of advertising (Gross Ratings Points, GRP) on sales, on purchases decision, and choices made. (Finding improved methods of measuring ROI is still an important area of research.) A classical introductory paper is due to Danaher and Rust [10]. Taylor [20] has summarized the current focus of research on advertising.

This is not the first paper that presents approaches based on graphs for recommendation systems. Huang et al. [13] proposed to build a bipartite graph of users and items, where user vertices are connected with item vertices if the user bought or gave a good evaluation to the item. The authors estimate the interest of a given user in a give item by aggregating the weights of short path between the user and the item in the graph. Lien and Phuong [16] extend the users-items bipartite graph with weights representing the evaluation the users gave to the items. Regarding flows, Adomavicius and Kwon [1] used a maximum flow algorithm for maximizing the diversity of the recommendations (instead of improving the recommendation accuracy as we do).

VI. CONCLUSIONS

We have presented a collection of graph-based recommendation engines, and have tested them in a real ecosystem of smartphone apps. The results obtained drive us to conjecture that using graphs for recommendation is a promising line of research. However, more experiments are needed in order to verify or disprove this conjecture.

In this work we have built recommendation engines that used graphs of items. We believe that graphs of users could also be very useful for recommendation. However, these graphs tend to be must larger (of several million nodes in our real system, versus a few hundreds of item), and processing them requires using more powerful computational systems and developing scalable algorithms.

REFERENCES

[1] G. Adomavicius, Y. Kwon, "Maximizing Aggregate Recommendation Diversity: A Graph-Theoretic Approach," in *Workshop on Novelty and Diversity in Recommender Systems*, 2011.

[2] Amazon Web Services, Inc., *Amazon Elastic MapReduce*, 2014. [Online http://aws.amazon.com/elasticmapreduce/; accessed 6-August-2014].

[3] T. Ambler, J. H. Robers, "Assessing marketing performance: don't settle for a silver metric," *Journal of Marketing Management*, 24 (7,8), 733-750, 2008.

[4] C. Anderson, "The Long Tail: Why the Future of Business Is Selling Less of More," *Hyperion*, 2006.

[5] The Apache Software Foundation, *The Apache Hadoop project*, 2014. [Online http://hadoop.apache.org/; accessed 6-August-2014].

[6] The Apache Software Foundation, *Apache HBase*, 2014. [Online http://hbase.apache.org/; accessed 6-August-2014].

[7] The Apache Software Foundation, *Apache Pig*, 2014. [Online http://pig.apache.org/; accessed 6-August-2014].

[8] S. Brin, L. Page, "The Anatomy of a Large-Scale Hypertextual Web Search Engine," *Computer Networks* 30(1-7): 107-117, 1998.

[9] H. Cordobés, A. Fernández Anta, L. F. Chiroque, F. Pérez, T. Redondo, A. Santos, "Graph-based Techniques for Topic Classification of Tweets in Spanish," *IJIMAI* 2(5): 31-37, 2014.

[10] P. J. Danaher, R. T. Rust, "Determining the optimal return on investment for an advertising campaign," *European Journal of Operational Research*, 95:511-521, 1996.

[11] L. R. Ford Jr., D. R. Fulkerson, *Flows in Networks*, Princeton University Press, 1962.

[12] J. L. Herlocker, J. A. Konstan, L. G. Terveen, J. T. Riedl, "Evaluating collaborative filtering recommender systems," *ACM Transactions on Information Systems*, vol 22(1), pp. 5-53, 2004 .

[13] Z. Huang, H. Chen, D. Zeng, "Applying Associative Retrieval Techniques to Alleviate the Sparsity Problem in Collaborative Filtering", *ACM Transactions on Information Systems*, vol. 22(1), pp. 116–142, 2004.

[14] D. Jannach, M. Zanker, A. Felfernig, G. Friedrich, *Recommender Systems: An Introduction*, Cambridge, 2010.

[15] Y. Koren, R. M. Bell, "Advances in collaborative filtering," in *Recommender Systems Handbook*, Springer, pp. 145–186, 2011.

[16] D. T. Lien, N. D. Phuong, "Collaborative filtering with a graph-based similarity measure," in *International Conference on Computing, Management and Telecommunications (ComManTel)*, 2014.

[17] G. Linden, B. Smith, J. York, "Amazon.com recommendations: Item-to-item collaborative filtering," *IEEE Internet Computing*, 7(1), 76-80, 2003.

[18] G. Shani, A. Gunawardana, "Evaluating Recommendation Systems," in *Recommender Systems Handbook*, Springer, pp. 257-297, 2011.

[19] R.C. Sprinthall, Basic Statistical Analysis, 9th Edition, *Pearson Education Group*, 2011.

[20] C. R. Taylor, "Hot topics in advertising research," *International Journal of Advertising*, 32(1), 2013.

Luis F. Chiroque obtained his B.Sc. degree in Telematics Engineering from the Polytechnic University of Madrid. Currently, he is studying a Master's Degree in Mathematical Engineering at University Carlos III of Madrid and he is a PhD student at IMDEA Networks, where he previously worked in the SOCAM project. His main research interest areas are graph theory, social networks and big data.

Héctor Cordobés obtained his MSc in Telecommunications in 2003 from Universidad Carlos III de Madrid. From that year he worked for Motorola, participating in international projects such as Motorola Soft Switch, Motorola IMS Core Network, High-Availability Computing Platform or Kreatel IPTV Platform, as Senior System Architect and Developer. In 2013 he joined IMDEA Networks as a Research Engineer, working in Natural Language Processing and Data Analysis projects.

Antonio Fernández Anta is a Research Professor at IMDEA Networks. Previously he was a Full Professor at the Universidad Rey Juan Carlos (URJC) and was on the Faculty of the Universidad Politécnica de Madrid (UPM), where we received an award for his research productivity. He was a postdoc at MIT from 1995 to 1997. He has more than 20 years of research experience, with a productivity of more than 5 papers per year on average. He is Chair of the Steering Committee of DISC and has served in the TPC of numerous conferences and workshops. He is a senior member of the IEEE and the ACM. He received his M.Sc. and Ph.D. from the University of Louisiana in 1992 and 1994, respectively. He completed his undergraduate studies at the UPM, having received awards at the university and national level for his academic performance.

Rafael A. García Leiva has a Bachelors degree in Computer Science by the University of Córdoba (Spain), a Master degree in Computational Sciences by the Univesity of Amsterdam, and a Diploma of Advanced Studies in Telematics by the Universidad Autónoma de Madrid. He worked during four years in the University of Córdoba as research assistant and scientific programmer, in the areas of Systems and Networks, Geographical Information Systems and Remote Senging. He worked during three years in the Universidad Autónoma de Madrid as research engineer in the area of High Energy Physics. He worked during three years as R&D manager at Andago Ingeniería, coordinating the R&D activities of the company. In 2008 he founded his own company, Entropy Computational Services, where he worked during five years in the areas of Social Networks, Mobile Applications, and Quantitative Trading. In 2014 he joined to the Institute IMDEA Networks, as research engineer, working in the area of Big Data and Machine Learning.

Philippe Morere is a member of the Research Support team at Institute IMDEA Networks. He works as a Research Engineer under the supervision of Antonio Fernández Anta, Research Associate Professor at the Institute. Prior to his incorporation to IMDEA Networks, Philippe worked as an Internship Student for Ritsumeikan University (Japan), where he was researching on embedded vision systems for intelligent robotic control. He obtained a B.Sc. degree in Engineering Sciences from the engineering school Enseirb-Matmeca, University of Bordeaux 1 (France). A few years later, he went on to pursue a M.Sc. degree in Computer Sciences & Telecommunications also at Enseirb-Matmeca.

Lorenzo Ornella received in 2014 his M.Sc. Degree in Computer and Communication Networks Engineering from the Politecnico di Torino (Turin, Italy). He previously obtained a Laurea in Ingegneria Delle Telecomunicazioni. He collaborated with IMDEA Networks and Zed from November 2013 until May 2014.

Fernando Pérez García is Big Data & Advanced Analytics Services manager at Zed. He was before at IBM, as a Data Specialist, in the Business Analytics & Optimization service line, developing Business Intelligence and Data Mining projects for main Spanish banks and carriers. Fernando is an entrepreneur, Android and Graphs enthusiast. Creator of Nautka Apps, Fernando has developed applications with over 150K users, providing natural language processing technologies to final users. Fernando is the inventor of the US Patent 7289025: "Method and system for securing an electronic device", on October 2007. Fernando completed his undergraduate studies (Licenciado en Informática) at the Universidad Pontificia de Salamanca in 1997.

Agustín Santos, PhD in Computer Science from Universidad Rey Juan Carlos de Madrid. He currently works at IMDEA Networks in several R+D projects. He previously developed his professional activities in the private sector. His broad experience covers technological base companies for which he has been founding partner. He has also held the positions of manager and chief director. His interests are focused in the fields of Distributed Systems, Big Data, Simulation, and Natural Language Processing.

Procedural Content Generation for Real-Time Strategy Games

Raúl Lara-Cabrera1, Mariela Nogueira-Collazo, Carlos Cotta and Antonio J. Fernández-Leiva

Lenguajes y Ciencias de la Computación department, Universidad de Málaga

***Abstract* — Videogames are one of the most important and profitable sectors in the industry of entertainment. Nowadays, the creation of a videogame is often a large-scale endeavor and bears many similarities with, e.g., movie production. On the central tasks in the development of a videogame is content generation, namely the definition of maps, terrains, non-player characters (NPCs) and other graphical, musical and AI-related components of the game. Such generation is costly due to its complexity, the great amount of work required and the need of specialized manpower. Hence the relevance of optimizing the process and alleviating costs. In this sense, procedural content generation (PCG) comes in handy as a means of reducing costs by using algorithmic techniques to automatically generate some game contents. PCG also provides advantages in terms of player experience since the contents generated are typically not fixed but can vary in different playing sessions, and can even adapt to the player herself. For this purpose, the underlying algorithmic technique used for PCG must be also flexible and adaptable. This is the case of computational intelligence in general and evolutionary algorithms in particular. In this work we shall provide an overview of the use of evolutionary intelligence for PCG, with special emphasis on its use within the context of real-time strategy games. We shall show how these techniques can address both playability and aesthetics, as well as improving the game AI.**

***Keywords* — Procedural Content Generation, Artificial Intelligence, game strategy, self-learning.**

I. Introduction

SPURRED on by the emergence of the videogame industry as the main component of the entertainment industry has motivated, research on videogames has acquired increasing notoriety during the last years. Such research spans many areas such as marketing and gamification, psychology and player satisfaction, computational intelligence, education and health (serious games) and computer graphics, just to cite a few. This diversification of research areas is largely motivated by a shift in the priorities of the video game industry: while games used to rely heavily on their graphical quality, other features such as the music, the player immersion into the game and interesting storyline have gained enormous importance. To cope with the plethora of new interesting challenges in the area of videogames, artificial and computational intelligence are turning out to be instrumental tools [25].

We recently carried out a mathematical, network-based study of the research community in the field of computational intelligence in video games [22] and obtained conclusive evidence of the vibrant activity of the field, which is steadily gaining momentum (as reflected in the growth patterns of new researchers and new publications). Still, the community of computational intelligence in video games is not yet fully developed, and collaboration links are still forming and improving the cohesion of the community. Besides, the industry is beginning to adopt the techniques and recommendations that academia offers.

Procedural Content Generation (PCG) refers to the algorithmic creation of content for video games, such as maps, levels, terrains, graphic textures, music, rules, quests, narrative, and missions among others possible [33]; traditionally, the creation of NPC (non-player controlled) behavior is not considered as PCG although, in a more global perspective, it is specific content for the game. The advantages of automatically creating video game content are manifold: firstly, it provides a drastic reduction in the cost and time of development as well as the memory used to store game artifacts; secondly, PCG provides a mechanism to inspire human artists to improve their creativity. Therefore, PCG can be considered from many different points of views and raises a high number of challenges from both Academic and Industry [35]. Moreover, the influence of PCG in, at least, other six areas in game programming, namely, NPC behavior learning, search and planning, games as Artificial Intelligence (AI) benchmarks, AI-assisted game design, general game AI, and AI in commercial games, underlines its importance [39, 40].

From the set of genres of videogames, Real-Time Strategy (RTS) games are one of the most exciting sub-genres since they require managing different kind of units and resources in real-time. In addition, they usually involve the participation of multiple players (not all of them necessarily human) that have to deal with incomplete information during the game; it is precisely this combination of resource management, multiplayer context and partial knowledge of the world what makes them an ideal framework to conduct Artificial Intelligence experiments; indeed, many challenging problems, such as resource allocation, adversarial real time planning,

DOI: 10.9781/ijimai.2015.325

spatial and temporal reasoning, opponent modeling, and opponent strategy prediction, just to name a few, can be addressed. As a result, RTS games offer a wide variety of fundamental AI research challenges [20].

In this context, one of the most interesting challenges in the videogame development process is precisely the procedural generation of content for RTS games as the artifact creation can be handled from many different perspectives due to the heterogeneity of the content that can be produced in RTS games, and to the participation of multiple (sometimes hundreds of) players with diverse profiles and skills. This work deals with the application of PCG techniques in RTS games, firstly by providing a brief review on this issue and, consequently, covering specific case studies in which evolutionary search has been employed to produce game components that satisfy certain properties.

II. PROCEDURAL CONTENT GENERATION

Videogames provide a wide range of fundamental problems that are useful for doing research in artificial intelligence. Among these we can cite real-time task planning and decision-making under uncertainty. This is particularly true in the case of Real-Time Strategy (RTS) games, which represent a whole genre of videogames in which the players must manage a collection of units and assets without a definite turn structure, that is, actions are asynchronously taken. Not surprisingly, RTS games have been used as researching tools to study and develop new artificial intelligence techniques, as explained in our paper about RTS games and computational intelligence [20].

The type of content that PCG techniques are able to create is very diverse, being maps and levels the prevailing type, as demonstrated by the large number of papers in the state of the art which are related to automatic level generation [14]. For example, Frade et al. introduced the use of genetic programming for evolving maps for video games, using in this process both human subjective evaluation and quality measures such as maps' accessibility [10] and edge length [11]. Another example of procedural level generation by Lanzi et al. [18] consists of evolving game maps that are specifically designed to improve the balance of the game, so no player has a marked superiority over the opponent (we will return to this issue later on).

Regarding other kind of content, Hastings et al. [12, 13], proposed a PCG algorithm for the game "Galactic Arms Race" in which the weapons available were generated on the fly. In this case, the fitness of the generated weapons was computed based on the amount of time the players used them, hence measuring the player satisfaction without requiring explicit feedback from the players. Onuczko et al. [31] presented a tool prototype for automatically producing specifications for missions and quests for a role-playing game. Font et al. [9] showed initial research towards a system capable of creating the rules for different card games. Collins [5] explored several approaches to procedural music composition.

Focusing on PCG for RTS games, Togelius et al. [36, 37] presented a multi-objective evolutionary algorithm whose objective was to create maps for this kind of games. Mahlmann et al. [26] described a search-based map generator for the game Dune 2, which was able to build playable maps using cellular automata (converting low-resolution matrices into maps fulfilling gameplay requirements). Finally, Ruela and Guimaraes [34] used a coevolutionary evolutionary algorithm aiming to maximize the performance of battle formations for the strategy game Call of Roma. We will also tackle coevolution later in this work.

III. CASE STUDIES

Historically, the success of a video game was directly associated with its graphical quality, but in the last decade this has changed and having good graphics does not necessarily ensure high sales. Players demand video games that show more than just a nice graphical quality and other issues, such as music, the story, or the atmosphere of the game, influence the decision of a player to get a specific game. The question of what it is that attracts the attention of players in a game is easy to answer: fun. How to obtain fun games and whether we can predict if the game will be of interest to players are not so easily answered, though.

There are several theories in the literature on what makes video games fun and why we play games [17], and, according to [4], a game's achievement might be deduced by measuring in advance the quality of the game (which seems however to be a difficult task). The notion of fun is difficult to measure as this depends on each player but it is naturally associated with the notion of player satisfaction: the greater the satisfaction, the greater the fun.

A. Playability-oriented PCG

This subsection focuses on the ability of PCG to engage the player (as commercial games demand) by keeping, during a match, an adequate trade off between the dynamism of the game and the balance between players which, probably, have different skills. More precisely, we aimed to generate maps for the RTS game Planet Wars, focusing on the properties that a priori make it entertaining and appealing to play, ensuring that the games are balanced (i.e., the forces of one of the players are not overwhelmingly larger than those of the other player – see [19]) and dynamic (i.e., action packed, there are battles and changes in the balance of power of the players – see [21]) For this purpose we are going to use evolutionary algorithms (EAs). An EA is a nature-inspired optimization and search method that deals with a set of entities (termed population), which represents a set of possible solutions. These entities, which are called individuals or chromosomes, compete against each other so the fittest individuals prevail over time, evolving towards better solutions. This is an iterative process where each step involves crossing (mixing information from several solutions) and mutating (performing random changes) individuals using genetic operators. Because individuals that represent the most appropriate solutions (as dictated by a so-

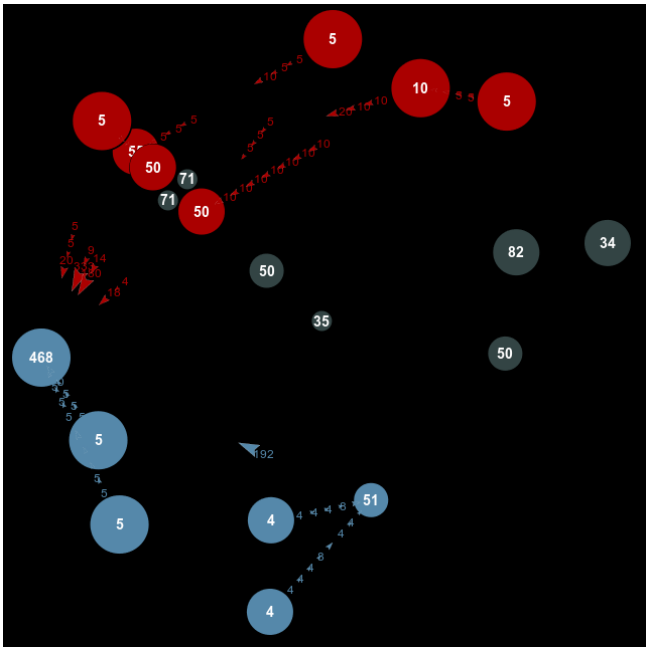


Fig. 1. A game of Planet Wars in progress. The arrows represent moving fleets while the number over the planets shows the number of stationed ships.

called fitness function that measure the goodness of solutions) are more likely to survive, the population gradually improves.

In order to use an EA, it is necessary to define several parameters: the individual's representation, the genetic operators, the size of the population and the number of generations the algorithm will be running.

Firstly, we had to consider how the solutions were to be represented and evaluated. A map for Planet Wars (see [24] for a description of the game) is defined as a collection of n_p planets distributed over a bi-dimensional plane. Each planet is characterized by its coordinates (x_i, y_i), its size s_i (determining the rate at which this planet produces new ships once captured by one of the players) and an initial number of ships w_i (determining the forces required to conquer the planet for the first time). As a result, a map can be described as a list $[\rho_1, \rho_2, \ldots, \rho_n]$, where each ρ_i is a tuple $\langle x_i, y_i, s_i, w_i \rangle$.

Two of the planets (the first two for simplicity) are initially marked as home planets of the players. From the point of view of the EA, the he number of planets n_p need not be fixed, and can range between an upper and a lower limit (15 and 30 in our experiments, see, e.g., Figure 1). In fact, one of the features of the evolutionary approach discussed later on is the ability to self-adapt to not only search parameters but also to the complexity (i.e., number of planets) of the map.

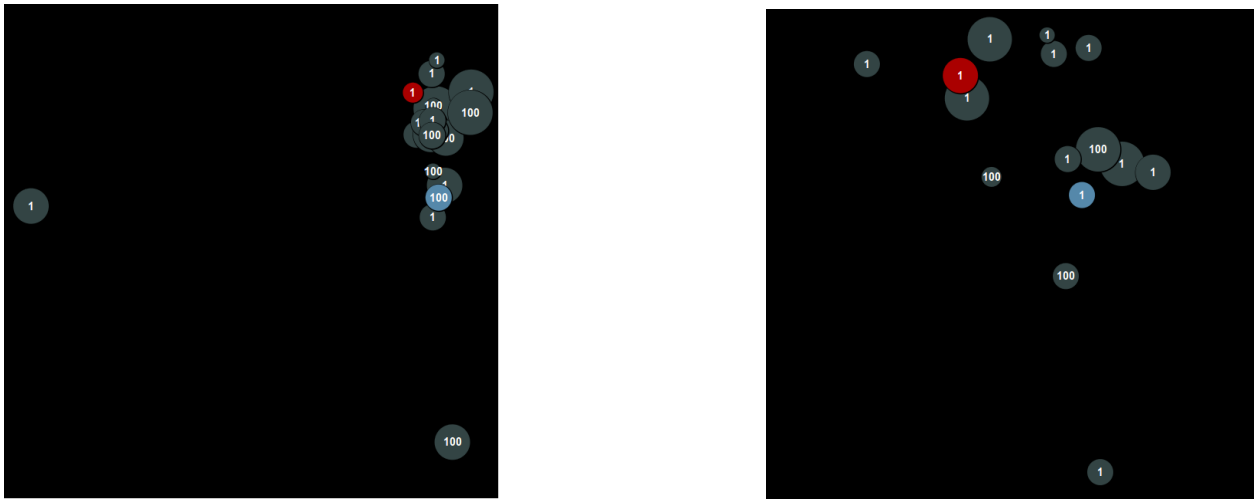
Regarding the evaluation of a map's playability features, we defined a tournament system which runs several games between an arbitrary number of pre-defined bots. Then, the tournament system analyzed some statistics gathered from each game in order to compute and quantify how balanced and dynamic the game was. Precisely, the system collects the following information from the *i*-th game (out of the total number of N_g games played in the tournament):

- *Territorial imbalance*: this is defined as the average imbalance in conquered planets throughout the game (the difference between the percentage of planets conquered by each player at each turn, averaged for all turns).
- *Growth imbalance:* this is measured analogously to the territorial imbalance, but considering the combined ship production capacity rather than the number of planets conquered (a player may have conquered many planets but these may be small, whereas other player may only dominated a few large planets).
- *Ship imbalance:* the same ideas sketched above are in this case applied to the number of ships (notice that a player can accumulate a large number of ships by following a passive strategy and not getting involved in fights and vice versa).
- *Game length:* this is just the percentage of the maximum number of turns played in the current game. Short games are imbalanced because it is implied that one of the player quickly destroys the fleet of its opponent.
- *Conquering rate:* this is the percentage of plantes conquered at the end of the game. If it is high, it means that the players have actively engaged in expanding their territories rather than sitting in their home planets.
- *Reconquering rate:* related to the previous measure, this is the average percentage of plantes whose ownership changes during the game (a high rate indicates that the players are actively fighting each other).
- *Peak difference:* this is actually a collection of variables, each of them measuring the maximal amplitude of the variation in any of the resources accounted for, in this case planets, combined size and ships.

These variables are subsequently averaged across the N_g games comprised in the tournament. In order to evaluate the actual balance and dynamism of a map we define a fuzzy rule base that captures some expert characterization of these features. For example, in order to account for balance we can use:

1) **if** territorial imbalance **is** LOW **and** growth imbalance **is** LOW **then** balance is HIGH
2) **if** territorial imbalance **is** HIGH **and** growth imbalance **is** low and ship imbalance **is** LOW **then** balance *is* MEDIUM
3) **if** (territorial imbalance **is** LOW **and** growth imbalance **is** HIGH) or game length **is** LOW **then** balance **is** LOW

Intuitively, we consider that a map has high balance if the average imbalance in planets and growth is low during the game. If one of the players has material advantage in terms of planets and ships, even if the combined growth is similar, the we deem the map to have medium balance. Finally, if both players manage to conquer a similar number of planets but their sizes are disparate or the game length is short, the map has low balance. Of course, there is some room for refining this characterization of balance by considering other combinations of the variables, but the above serves as an illustrative example. Similarly, we can define fuzzy rules for dynamism. For example, we can state that



(a) (b)

Fig. 2. Two examples of maps that have been generated by the algorithm. Planet's colors denote whether it is conquered by some player (red/blue) or remains neutral. The number shows how many ships are defending each planet.

1) **if** conquering rate **is** HIGH **and** reconquering rate **is** HIGH **then** *dyn* **is** HIGH
2) **if** all peak differences **is** HIGH **then** *dyn* **is** HIGH

i.e., if there are many planets being conquered and reconquered, or the peak differences in all three resources is high, then dynamism is high (there are battles and action).

Likewise if it turns out that one of the peak differences is high but any of the other two is not, the dynamism can be said to be intermediate (this would be captured in a family of three rules). Finally, if all peak differences are low or the conquering and reconquering rates are not high and the game length is very short, the dynamism of the map is considered to be low.

The procedural map generator used a self-adaptive evolutionary approach with the solutions encoded as mixed real-integer vectors. The parameters governing mutation were also a part of the solutions, thus providing the means for self-adapting them (see [24] for a full explanation of the evolutionary algorithm and its parameters and operators). The players of the tournament system used to assess the quality of the maps during the evaluation phase were three bots submitted to the *Google AI Challenge 2010*, namely *Manwe*, *Flagscapper's bot* and *fglider's bot*. All of them ranked in the top 100 (there were over 4600 participants) and their source code was available – see [24] for the URLs.

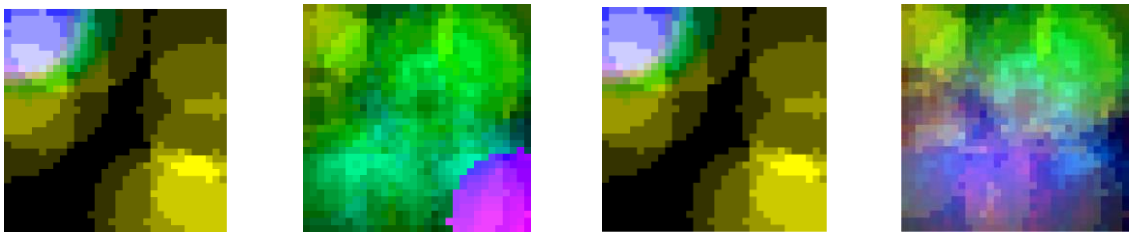
Experiments focusing separately in either of the two properties point at the higher difficulty of attaining dynamism with respect to balance. Figure 2 shows an example of the maps obtained[6]. Of course it is possible to optimize both properties at the same time following a multi-objective approach (the Non-dominated Sorting Genetic Algorithm II – *NSGA-II*– in our case). By doing so, we can obtain a collection of solutions representing different tradeoffs between balance and dynamism (ranging from highly balanced and lowly dynamic to highly dynamic and poorly balanced, with different intermediate scenarios in which an increase in one the properties is traded by a decrease in the other). Note in this sense that a single-objective approach cans easily exploit the first objective (i.e. balance), providing maps that achieve perfect balance due to the complete inaction of the players. However, the situation is different from the point of view of dynamism, since according to our definition a very unbalanced game is likely going to be short or feature less alternation between the players, hence resulting to be non-dynamic as well. For this reason, the multiobjective approach yields a graceful degradation of dynamism when balance is increased, eventually exhibiting an abrupt reduction of the dynamism upon reaching the high end of balance. Further studies show that, in general, dynamic games seem to be related to maps featuring a larger number of planets, widely scattered on the map and whose sizes are positively correlated to the initial number of ships.

[6] It is possible to watch a game on these maps at http://www.lcc.uma.es/~raul/maps/maps.html

B. Introducing Aesthetics

In Section III-A we focused on making the game more fun to play, obtaining games that are balanced and dynamic. However, the generated maps lacked aesthetics (for example, maps with all their planets clustered in a small region, see Figure 2), which is an interesting feature apart from the fun that may lead to increase the player satisfaction. It turns out that fun and aesthetics are two complementary ways of achieving the same goal [27]. Moreover, non- aesthetic maps may confuse the player, reducing his/her satisfaction or even leading him/her to stop playing the game.

Following a similar evolutionary scheme and representation of the solutions for the automatic generation of balanced and dynamic maps, we considered different properties in order to evaluate the aesthetics of maps. We establish a separation between geometrical features (based on the spatial properties of the map, namely coordinates and distances), and topological features (based on qualitative relationships among planets invariant under geometrical



(a) (b) (c) (d)

Fig. 3. Map's distribution over the SOM for both geometric (a) and topological (b) approaches. Yellow for non-aesthetic, cyan for aesthetic and magenta for non-dominated. (c) and (d) show the topological approach solution projected over the geometric approach SOM and vice versa.

transformations such as rotation, translation or scaling). We also take into account morphological features based on individual planet properties, such as size or initial number of ships.

These are the geometrical measures:

- *Spatial distribution of planets*: given planet coordinates we compute the average distance between planets μ_d and the standard deviation of these distances σ_d.
- *Planet features*: given the sizes and initial number of ships of each planet, we compute the average and standard deviation of sizes (μ_s and σ_s respectively) and Pearson's correlation ϱ between sizes and number of ships.

Thus, we can characterize a map by a 5-tuple $\langle \mu_d, \sigma_d, \mu_s, \sigma_s, \varrho \rangle$, and use some distance measure (e.g., Euclidean distance) to determine the geometrical distance among two maps.

As to the topological features, these are extracted from the sphere-of-influence graph (SIG) of each map, which sets a relationship between some set of points based on their spatial arrangement [38] (defining a planet's radius of influence as the shortest distance of any other planet, and defining a graph in which vertex is a planet and edges are defined between planets whose distance is less or equal to the sum of their respective radii of influence). Using this SIG we can compute:

- *Number of connected components*: number of maximal sub-graphs in which any two vertices are connected by at least one path.
- *Average node's degree*: average number of edges incident to each node.
- *Density of the graph*: ratio between the number of edges of the graph and that of a complete graph with the same number of vertices.
- *Average clustering coefficient*: average percentage of each node's neighbors which are neighbors of each other too.
- *Pearson correlation between the size of the nodes and their betweenness centrality*. Betweenness is a measure of the importance of each node as an intermediate gateway in the paths between any other two nodes. We measure is highly central nodes are also large planets.
- *Pearson correlation between the size of the nodes and their degree*.
- *Size assortativity*, i.e., Pearson correlation coefficient between the size of nodes connected in the graph (i.e., the extent to which plantes are linked to other planets of larger or smaller size)

As with geometrical measures, these topological measures can be used to characterize a map and define a distance metric among them. However, some of these measures turn out to be somewhat redundant. By considering a collection of 20 maps (10 with good aesthetics and 10 with bad aesthetics as tagged by a human expert) and using a Random Forest classifier to determine which measures are useful for classification purposes we obtain that graph's density, correlation between node size and betweenness and size assortativity are the most relevant ones – see [23] for further details.

If we run an EA using distance to aesthetic maps (to be minimized) and to non-aesthetic maps (to be maximized) in a multi-objective approach, we observe that there is a smooth, linear transition between these two objectives. More qualitatively, we created two self-organizing map (SOM) [16] with 32×32 process units over a non-toroidal rectangular layout, one for each characterization approach (geometrical and topological). As we can see in Figure 3, the SOM of the geometrical approach set a separation between non-aesthetic (yellow zones) and aesthetic maps (cyan zones), as well as generated maps (magenta zones) share the same region as aesthetic maps. Thus, they can be considered aesthetic as well. Regarding the topological approach, the distinction between aesthetic and non-aesthetic maps is not so clear though, as shown by the overlapped areas.

C. Self-learning of RTS strategies

As another branch of PCG, the search of game strategies via computational intelligence (CI) emerges as an important sub-field. RTS games are specifically distinguished for imposing the players the control of many different resources during the game. For this reason the procedural generation of game strategies should be backed up by methods allowing a significant reduction of the computational time involved in the exploration of the large search spaces implied. We are here specifically concerned with the use of techniques providing continuous, autonomous learning capabilities for the artificial intelligence embedded in a RTS game. We consider coevolution for this purpose.

Coevolution is a model inspired in the principles of natural evolutionary theory. It is based on the interaction between different species and can take two forms: one based in the collaboration and other one based on competition. Cooperative approaches simulate a symbiotic relationship, used for finding

Fig. 4. Screenshot of RobotWars game.

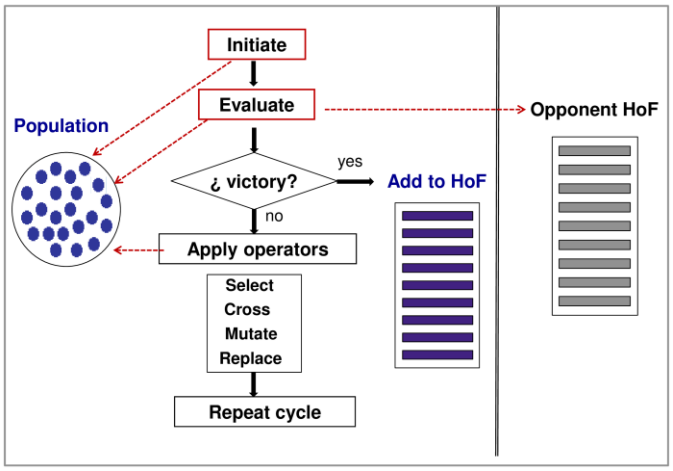


Fig. 5. Basic coevolutionary cycle which uses the HoF during the evaluation process.

a solution through the collaboration between many possible solution components; on the other hand, competitive approaches establish a competition between individuals much like a predator/prey environment. The goal is to trigger an "arms race" in which the improvement of some individuals stimulates the improvement in the opponents, and vice versa. This last approach is usually used for solving optimization problems in inherently competitive contexts like games.

Several experiments have showed significant results in the application of coevolutionary models as a mechanism of self-learning in a RTS. For example, different variants of competitive coevolutionary (CC) algorithms [3], [15], [1] have been proposed to find optimal strategies for the Tempo game. Also, the authors of [2] analyzed the employment of coevolution for creating a tactical controller for small groups of game entities in a real-time capture-the-flag game. The proposal described in [6] explores several methods for automatically shaping the coevolutionary process by modifying the fitness function as well as the environment during evolution.

The success of the application of coevolutionary approaches is out of question but coevolution has also its own intrinsic problems – see [8], [7]. In particular the evaluation mechanism is a key point in a coevolutionary model because it guides the arms race that emerges from the interactions between individuals. For this reason several evaluation approaches have been proposed in the literature to alleviate some of the coevolutionary pathologies. In this line of work we have already explored the use of the Hall-of-Fame (HoF) [32] based mechanism as an archive method to memorize the successful solutions to guide the search process for generating game strategies in RTS games. This mechanism is used to provide a long term memory of the coevolutionary process, avoiding that some good strategies are forgotten due to lack of selective pressure.

Our first works [28],[29] were conducted in the context of the RTS game *RobotWars*[7]. The main goal was generating game strategies to control the behavior of an army. RobotWars is a self-developed game for testing game AI strategies (i.e. bots), and hence human players do not have place here. It is a two player's game, in which two different armies fight in a 3D scenario with many obstacles (Figure 4 shows a screenshot of this game). Each army has different units and one general; if an army wipes out the enemy general them they will be the winners of the game.

Using this RTS game five variants of a CC algorithm using HoF as a memory mechanism to keep the winning strategies were tested. In our model the individual was represented as a matrix of actions that allows to control, deterministically, the behavior of an army during the game. The basic coevolutionary schema implemented is showed in Figure 5. It is based in coevolutionary turns of multiple strategies for each army. The goal is to find a winning strategy which is then put in that player's HoF. That HoF is then used in the evolution of strategies for the other army until a new winning strategy is found, placed in the corresponding HoF, and the roles are reversed again. If at the end of the coevolutionary turn no solution is obtained, a new turn starts again until a champion is found or until the maximum number of cycles is reached.

During experiments in RobotWars we analyzed how the diversity and growth of the HoF can influence the quality of the solutions obtained by HoF-based CC algorithms. In this sense we studied the performance of eleven algorithms based on different mechanisms for maintaining and updating the champions' memory during the evaluation process. This was aimed to reduce the size of the HoF (hence reducing computational time) but doing so in an intelligent way, without losing the beneficial contribution of the long term memory. A *diversity* indicator based on the contribution to each champion to the diversity of the HoF diversity showed a good performance (i.e., the HoF was reducing by removing similar champions which did not contribute much to the coevolutionary learning). We also detected that manipulating the size of the HoF has a direct influence on the quality of the search result due to the loss of transitivity (a solution *A* beating another solution *B* which in turns beats *C* which can however beat *A*), so this should be done carefully.

That previous work was extended in [30] proposing a different evaluation mechanism to exploit the potential offered

[7] http://www.lcc.uma.es/~afdez/robotWars

by archive methods to maintain transitivity between the solutions; we considered a new RTS game –*Planet Wars*, described before– allowing a deeper experimental analysis and more consistent conclusions. This time we added novel strength indicators that were independent from the fitness function with the objective of avoiding the appearance of cycling (strategies being forgotten and re-discovered over and over again). The novelty of this last aspect consisted of incorporating into our prime CC algorithm which used the HoF as shown in Figure 5, an additional archive (termed call-of-celebrities, HoC) that contained a team of experienced virtual players. These were used to evaluate how strong a candidate was. The combined use of both halls (HoF and HoC) with the (possibly combined) utilization of diversity and quality metrics helped the optimization to obtain competitive bots that self-adapt to beat their (co)evolved enemies.

IV. CONCLUSION

Procedural Content Generation (PCG) is one of the corner stones of the modern video game industry. Throughout this paper we have described three case studies that are part of our work in the area of PCG for real-time strategy video games. In the first place, we have presented and compared several methods for generating maps for the game Planet Wars; such maps are firstly oriented to fulfill the requirements of the player in terms of playability, that is, providing an interesting and enjoyable experience as to what the game mechanics regards. This has been done characterizing some positive features a game should have such as balance (having an opponent with similar skills as the player, as reflected in the achievements of the former in the game with respect to those of the latter) and dynamism (delivering an existing game in which numerous events unfold and there are changes in the balance of power between the two players). It has been shown how maps with these features can be accomplished by using an evolutionary approach for their automatic generation. Subsequently, we have considered the aesthetics perspective. Given the highly subjective nature of this endeavor, the input of an expert is required in order to provide samples of aesthetic/non-aesthetic maps, which can be in turn used by an evolutionary algorithm as reference to reproduce features of good maps, and avoid features of bad maps. Such features admit different characterizations; we have described the use of both geometrical (based on the spatial distribution of map components), morphological (based on the individual properties of map components) and topological (based on properties of the maps which are invariant under simple geometrical transformations). By using an unsupervised learning method we can infer that an evolutionary approach based on these characterizations is capable of producing aesthetic maps.

Afterwards, we have extended the classical view of PCG by considering game AI as game content; in particular, we have considered NPC behavior and we have briefly described a self-learning approach that we employed on two RTS games with significant success. To do so, we used co-evolutionary techniques to lead the search process in a competitive context; we have also shown that our algorithmic proposals were based on the concept of Hall-of-fame (HoF) that basically represents a memory that allows to store the best candidates that are further employed in the evaluation phases to improve the optimization process. A number of different structures and mechanisms to select the champions to be stored in the HoF can be defined and this selection can have drastic influence in the results.

Many lines remain open; for instance, in order to accelerate the creation process (and as consequence, to minimize development costs), the industry demands the automatic generation of diverse content at the same time; moreover, there are artifacts that surely influence the creation of other class of elements, and vice versa. This basically means that PCG should be defined to enable the generation of contents (of distinct nature) at the same time with the goal of producing compound components. Our next step follows precisely this line of research and it consists of designing PCG methods to co-evolve graphical content (e.g., maps/levels) and game AI. In addition, obtaining correct quality metrics is an area that deserves more research; the evolutionary search directed to find high quality content heavily depends on the fitness functions that guide the optimization process, and it is not easy to evaluate the goodness of these; moreover, content creation is directly related to human creativity and, therefore, humans (both developers and players) are required to be involved in the evolution process: in this sense, designing correct user-centric interaction evolutionary models is also another line of exciting research.

REFERENCES

[1] P. Avery et al., "Coevolving a computer player for resource allocation games: using the game of tempo as a test space." Ph.D. dissertation, School of Computer Science University of Adelaide, 2008.

[2] [2] P. Avery and S. J. Louis, "Coevolving team tactics for a real-time strategy game," in IEEE Congress on Evolutionary Computation. Barcelona, Spain: IEEE, 2010, pp. 1–8.

[3] P. M. Avery and Z. Michalewicz, "Static experts and dynamic enemies in coevolutionary games," in IEEE Congress on Evolutionary Computation. IEEE, 2007, pp. 4035–4042.

[4] C. Browne and F. Maire, "Evolutionary game design," IEEE Trans. Comput. Intellig. and AI in Games, vol. 2, no. 1, pp. 1–16, 2010.

[5] [5] K. Collins, "An introduction to procedural music in video games," Contemporary Music Review, vol. 28, no. 1, pp. 5–15, 2009.

[6] A. Dziuk and R. Miikkulainen, "Creating intelligent agents through shaping of coevolution," in IEEE Congress on Evolutionary Computation. New Orleans, LA, USA: IEEE, 2011, pp. 1077–1083.

[7] M. Ebner, R. A. Watson, and J. Alexander, "Coevolutionary Dynamics of Interacting Species," in Applications of Evolutionary Computation, EvoApplicatons 2010 (EvoApplications (1)), ser. Lecture Notes in Computer Science, C. D. Chio et al., Eds., vol. 6024. Istanbul, Turkey: Springer, 2010, pp. 1–10.

[8] S. G. Ficici and A. Bucci, "Advanced tutorial on coevolution," in 2007 GECCO Conference Companion on Genetic and Evolutionary Computation. New York, USA: ACM, 2007, pp. 3172–3204.

[9] J. Font, T. Mahlmann, D. Manrique, and J. Togelius, "A card game description language," in Applications of Evolutionary Computation, ser. Lecture Notes in Computer Science, A. Esparcia-Alcázar, Ed. Springer Berlin Heidelberg, 2013, vol. 7835, pp. 254–263.

[10] M. Frade, F. F. de Vega, and C. Cotta, "Breeding terrains with genetic terrain programming: The evolution of terrain generators," International Journal of Computer Games Technology, vol. 2009.
[11] ------, "Evolution of artificial terrains for video games based on obstacles edge length," in IEEE Congress on Evolutionary Computation. IEEE, 2010, pp. 1–8.
[12] E. J. Hastings, R. K. Guha, and K. O. Stanley, "Automatic content generation in the Galactic Arms Race video game," Computational Intelligence and AI in Games, IEEE Transactions on, vol. 1, no. 4, pp. 245–263, 2009.
[13] ------, "Evolving content in the Galactic Arms Race video game," in Computational Intelligence and Games, 2009. CIG 2009. IEEE Symposium on. IEEE, 2009, pp. 241–248.
[14] M. Hendrikx, S. Meijer, J. Van Der Velden, and A. Iosup, "Procedural content generation for games: A survey," ACM Trans. Multimedia Comput. Commun. Appl., vol. 9, no. 1, pp. 1:1–1:22, Feb. 2013.
[15] R. Johnson, M. Melich, Z. Michalewicz, and M. Schmidt, "Coevolutionary Tempo game," in Evolutionary Computation. CEC'04. Congress on, vol. 2, 2004, pp. 1610–1617.
[16] T. Kohonen, "The self-organizing map," Proceedings of the IEEE, vol. 78, no. 9, pp. 1464–1480, 1990.
[17] R. Koster, A theory of fun for game design. Paraglyph Press, 2004.
[18] P. L. Lanzi, D. Loiacono, and R. Stucchi, "Evolving maps for match balancing in first person shooters," in 2014 IEEE Conference on Computational Intelligence and Games, CIG 2014, Dortmund, Germany, August 26-29, 2014, 2014, pp. 1–8.
[19] R. Lara-Cabrera, C. Cotta, and A. J. Fernández-Leiva, "A procedural balanced map generator with self-adaptive complexity for the real-time strategy game planet wars," in Applications of Evolutionary Computation 2013, ser. Lecture Notes in Computer Science, A. Esparcia-Alcázar et al., Eds., vol. 7835.Berlin Heidelberg: Springer-Verlag, 2013, pp. 274–283.
[20] ------, "A review of computational intelligence in RTS games," in IEEE Symposium on Foundations of Computational Intelligence, FOCI 2013, Singapore, Singapore, April 16-19, 2013, 2013, pp. 114–121.
[21] ------, "Using self-adaptive evolutionary algorithms to evolve dynamism-oriented maps for a real time strategy game," in Large-Scale Scientific Computing - 9th International Conference, LSSC 2013, Sozopol, Bulgaria, June 3-7, 2013. Revised Selected Papers, 2013, pp. 256–263.
[22] ------, "An analysis of the structure and evolution of the scientific collaboration network of computer intelligence in games," Physica A: Statistical Mechanics and its Applications, vol. 395, no. 0, pp. 523 – 536, 2014.
[23] ------, "Geometrical vs topological measures for the evolution of aesthetic maps in a rts game," Entertainment Computing, vol. 5, no. 4, pp. 251–258, 2014.
[24] ------, "On balance and dynamism in procedural content generation with self-adaptive evolutionary algorithms," Natural Computing, vol. 13, no. 2, pp. 157–168, 2014.
[25] S. M. Lucas, M. Mateas, M. Preuss, P. Spronck, and J. Togelius, Eds., Artificial and Computational Intelligence in Games, ser. Dagstuhl Follow-Ups, vol. 6. Schloss Dagstuhl - Leibniz-Zentrum fuer Informatik, 2013.
[26] T. Mahlmann, J. Togelius, and G. N. Yannakakis, "Spicing up map generation," in EvoApplications, ser. Lecture Notes in Computer Science, C. D. Chio et al., Eds., vol. 7248. Springer, 2012, pp. 224–233.
[27] M. Nogueira, C. Cotta, and A. J. Fernández-Leiva, "On modeling, evaluating and increasing players' satisfaction quantitatively: Steps towards a taxonomy," in Applications of Evolutionary Computation, ser. Lecture Notes in Computer Science, C. D. Chio et al., Eds., vol. 7248.Málaga, Spain: Springer-Verlag, 2012, pp. 245–254.
[28] M. Nogueira, J. Gálvez, C. Cotta, and A. J. Fernández-Leiva, "Hall of Fame based competitive coevolutionary algorithms for optimizing opponent strategies in a new RTS game," in 13th annual European conference on simulation and AI in computer games (GAME-ON 2012), A. F.-L. et al., Ed. Málaga, Spain: Eurosis, November 2012, pp. 71–78.
[29] M. Nogueira Collazo, C. Cotta, and A. J. Fernández Leiva, "An analysis of hall-of-fame strategies in competitive coevolutionary algorithms for self-learning in RTS games," in Learning and Intelligent Optimization - 7th International Conference, LION 7, Catania, Italy, January 7-11, Revised Selected Papers, ser. Lecture Notes in Computer Science, G. Nicosia and P. M. Pardalos, Eds., vol. 7997. Springer, 2013, pp. 174–188.
[30] ------, "Virtual player design using self-learning via competitive coevolutionary algorithms," Natural Computing, vol. 13, no. 2, pp. 131–144, 2014.
[31] C. Onuczko, D. Szafron, J. Schaeffer, M. Cutumisu, J. Siegel, K. Waugh, and A. Schumacher, "Automatic story generation for computer role-playing games." in AIIDE, 2006, pp. 147–148.
[32] C. Rosin and R. Belew, "New methods for competitive coevolution," Evolutionary Computation, vol. 5, no. 1, pp. 1–29, 1997.
[33] N. Shaker, J. Togelius, and M. J. Nelson, Procedural Content Generation in Games: A Textbook and an Overview of Current Research. Springer, 2014.
[34] A. Siqueira Ruela and F. Gadelha Guimaraes, "Coevolutionary procedural generation of battle formations in massively multiplayer online strategy games," in Computer Games and Digital Entertainment (SBGAMES), 2014 Brazilian Symposium on, Nov 2014, pp. 89–98.
[35] J. Togelius, A. J. Champandard, P. L. Lanzi, M. Mateas, A. Paiva, M. Preuss, and K. O. Stanley, "Procedural content generation: Goals, challenges and actionable steps," in Artificial and Computational Intelligence in Games, ser. Dagstuhl Follow-Ups, S. M. Lucas, et al., Eds. Schloss Dagstuhl - Leibniz-Zentrum fuer Informatik, 2013, vol. 6, pp. 61–75. [Online].
[36] J. Togelius, M. Preuss, N. Beume, S. Wessing, J. Hagelback, and G. N. Yannakakis, "Multiobjective exploration of the Starcraft map space," in Computational Intelligence and Games (CIG), 2010 IEEE Symposium on. IEEE, 2010, pp. 265–272.
[37] J. Togelius, M. Preuss, and G. N. Yannakakis, "Towards multiobjective procedural map generation," in Proceedings of the 2010 Workshop on Procedural Content Generation in Games. ACM, 2010, p. 3.
[38] G. T. Toussaint, "A graph-theoretic primal sketch," Computational Morphology, pp. 229–260, 1988.
[39] G. Yannakakis and J. Togelius, "A panorama of artificial and computational intelligence in games," Computational Intelligence and AI in Games, IEEE Transactions on, vol. PP, no. 99, pp. 1–1, 2014.
[40] Broncano, C. J., C. Pinilla, R. Gonzalez-Crespo, and A. Castillo-Sanz, "Relative Radiometric Normalization of Multitemporal images", International Journal of Artificial Intelligence and Interactive Multimedia, vol. 1, issue A Direct Path to Intelligent Tools, no. 3, pp. 53-58, 12/2010

Raúl Lara-Cabrera received the Bachelor's degree in computer science and the Master's degree in software engineering and artificial intelligence from the University of Málaga (UMA), Spain, where he is currently pursuing the Ph.D., focusing his research on fields related to artificial intelligence and video games. He is currently a researcher with the Lenguajes y Ciencias de la Computación department.

Mariela Nogueira-Collazo obtained the Bachelors degree in Computer Science by the University of Computer Science (UCI) of The Havana, Cuba; and the Master's degree in Software Engineering and Artificial Intelligence from the University of Málaga (UMA), Spain. Currently she is pursuing the Ph.D. in UMA and her research cover the application of artificial intelligence in videogames context.

Carlos Cotta obtained his MSc and PhD in Computer Science from the University of Málaga (UMA), Spain in 1994 and 1998 respectively. He holds a tenured Professorship in the Department of Lenguajes y Ciencias de la Computación of this University since 2001. His main research areas involve metaheuristic optimization -in particular hybrid and memetic approaches- with a focus on both algorithmic and applied aspects (particularly combinatorial optimization) as well as complex systems.

Antonio J. Fernández-Leiva received, in 2002, the PhD degree in Computer Science from the University of Málaga (UMA), where he is currently associate professor in the Lenguajes y Ciencias de la Computación department. In the past, he worked in private companies as computer engineer. His main areas of research involve both the application of metaheuristics techniques to combinatorial optimization and the employment of Computational Intelligence in Games. He also leads a Master on Design and Programming of Videogames at UMA.

Mining Web-based Educational Systems to Predict Student Learning Achievements

José del Campo-Ávila, Ricardo Conejo, Francisco Triguero, Rafael Morales-Bueno

Universidad de Málaga, Andalucía Tech, Departamento de Lenguajes y Ciencias de la Computación, Campus de Teatinos, Málaga, España

***Abstract* — Educational Data Mining (EDM) is getting great importance as a new interdisciplinary research field related to some other areas. It is directly connected with Web-based Educational Systems (WBES) and Data Mining (DM, a fundamental part of Knowledge Discovery in Databases).**

The former defines the context: WBES store and manage huge amounts of data. Such data are increasingly growing and they contain hidden knowledge that could be very useful to the users (both teachers and students). It is desirable to identify such knowledge in the form of models, patterns or any other representation schema that allows a better exploitation of the system. The latter reveals itself as the tool to achieve such discovering. Data mining must afford very complex and different situations to reach quality solutions. Therefore, data mining is a research field where many advances are being done to accommodate and solve emerging problems. For this purpose, many techniques are usually considered.

In this paper we study how data mining can be used to induce student models from the data acquired by a specific Web-based tool for adaptive testing, called SIETTE. Concretely we have used top down induction decision trees algorithms to extract the patterns because these models, decision trees, are easily understandable. In addition, the conducted validation processes have assured high quality models.

***Keywords* — Data Mining, Decision Trees, Educational technology, Knowledge discovery.**

I. Introduction

SINCE Internet opened a new way to communicate in many different forms, the educational sector adopted such technology and developed the Web-based Educational Systems (WBES). Firstly, they were static systems, mainly dedicated to divulgate contents. But progressively, they extended their capabilities with new characteristics in order to make the systems adaptive and intelligent [1].

At this moment there exist many different systems that combine different elements to achieve some level of intelligence. Therefore, we can find WBES with adaptive techniques [2], some other WBES with intelligent mechanisms [3] and more complex systems that combine both properties (a detailed review of AIWBES was presented by Brusilovsky and Peylo [4]).

What it is evident is the high volume of data that these systems are storing and processing continuously: relations between contents offered to students, interactions with students, number of visits, marks achieved in tests, time used to respond those tests, etc.

Knowledge discovery in databases (KDD) continues extending to almost every field where large amount of data are stored and processed (databases, system logs, activity logs, etc.), so WBES becomes another environment to apply KDD processes.

The data mining techniques are essential for one of the most important points of KDD: they are applied in data analysis phase and machine learning algorithms are used to produce the models that summarize the knowledge discovered [5]. Therefore, it is easy to see that educational tasks can benefit from the knowledge extracted by data mining.

This research field is called Educational Data Mining (EDM) and its main objective is to analyze data stored in WBES in order to resolve educational research issues [6]: validation of the educational system, prediction of students learning achievements, identification of misconceptions [7], assessment and feedback to the authors of courses [8], etc.

In this paper we try to determine that data mining techniques can help to predict students learning achievements, mainly oriented to find relations between continual assessment (or evaluation) and the final grade achieved.

This paper is organized as follow. In Section 2 we describe the materials used and the conducted methodology. Basically, our materials are data collected by SIETTE[8], a Web-based tool for adaptive testing [9] and the framework for data mining called Weka [10]. Then, in Section 3, we present the results and comment the patterns discovered by machine learning algorithms. Finally, in Section 4, we summarize the most relevant conclusions and propose new research lines for futures works.

II. Materials and Methods

Considering the features offered by data mining in order to discover patterns in datasets, in this case extracted from Web-based Educational System, we propose to study the existence of different kinds of relations between the continuous

[8] http://www.siette.org

DOI: 10.9781/ijimai.2015.326

evaluation of students and their final achievements in the subject. For this purpose we work with the following materials and methodologies.

A. Materials

The raw materials of any process of knowledge discovery that uses data mining techniques are data, grouped in subsets called datasets. Every dataset is composed of examples described by attributes and labeled with a class (supervised learning). Values for these attributes can be numerical or nominal.

For this study we have focused in students that took the subject "Principles in Informatics" in two consecutive courses. The skills and competences to be achieved are varied: from basic concepts related with Computer Science (hardware, software, algorithms, etc.) to elementary abilities to develop computer programs using the C programming language.

The evaluation of this subject includes a continuous evaluation during the course (with a weight of 40% in the final grade) that ends with a final evaluation exam (60% weight). The continual assessment (or continuous evaluation) is compound of three tests (20%) and three practical exercises (20%). What we are using in this study are the marks achieved by the students in the tests that have been completed using the SIETTE Web-based Educational System [9]. First test (T1) is used to check how concepts related with Computer Science are assimilated, the second one (T2) focus on initial programming abilities with C (types, expressions, operators and control flow) and the third one (T3) check the knowledge about more advanced concepts in C (functions and structures). The final exam is mostly prepared to check the programming abilities; so basic concepts related with Computer Science are only evaluated with one test (T1).

In Table I we show some statistics related to the real marks achieved in the tests. The maximum value cannot be greater than 100.00, but minimum values can be lower than 0.00 because wrongly answered questions count negatively (if a student answers many questions incorrectly, the mark is lower than 0.00).

Taking this context in consideration, now we can describe the datasets that we have used. In our case the examples summarize the evaluation achieved by the students (116) that took the subject "Principles in Informatics". In a first approach we only consider the marks for every test, but in a second step we added the differences with respect to the average value, in order to establish a relative comparison between the results.

The class attribute is the final grade achieved in the global subject evaluation. We have used the numerical grade, defined in [0,10], and transformed it to the European ECTS grading scale (A for the best grades and F for the worst ones, F corresponds to students that fail) [11].

To carry out the mining process there exist different frameworks that implement multiple machine learning algorithms. We have used Weka [10] because it includes TDIDT (Top Down Induction Decision Trees) algorithms that represent the knowledge extracted in form of decision trees

TABLE I
MARKS ACHIEVED IN TESTS

	T1 (HW, SW, algorithms)	T2 (types, operators, control flow)	T3 (functions, structures)
Minimum	-20.00	-20.00	-18.33
Average	34.18 ± 19.30	35.50 ± 23.42	30.47 ± 27.84
Maximum	76.67	86.67	100.00

Minimum, average and maximum values observed in the tests answered by students. Maximum never can be greater than 100.00, but minimum values can be lower than 0.00 because wrongly answered questions count negatively

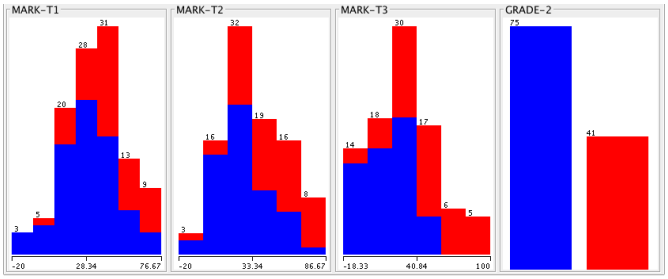


Fig. 1. Class distribution visualizing all marks (for tests T1, T2 and T3) in the first dataset (116 students). Blue color represents students that fail the evaluation (or absent themselves) and red color represents students that pass the evaluation. This chart is plotted by the Weka framework.

[12]: a model easily understandable by humans with some other additional advantages (learning with numerical or nominal data, robustness, verifiable reliability, etc.). Concretely we have selected the J48 algorithm (C4.5 [13] implementation coded in Weka), using it with its default configuration. When plotting the decision trees (Fig. 4, 5, and 6), the numbers present in the nodes (<first> / <second>) represent the number of examples that satisfy the branch (<first>) and the number of examples that, in addition, are incorrectly classified (<second> that it is not present when there is no errors).

B. Methods

Once we have described the datasets and the framework we have used, we can detail which methodology we have followed.

Firstly we have preprocessed the data in order to clean and prepare them. Data extracted from SIETTE are very rich and diverse, but nowadays, they cannot be directly exported to the kind of dataset supported by Weka (ARFF files). Some transformation steps were needed: discretization of numerical grade to ECTS grading scale, calculation of new calculated attributes, identification of missing values, etc.

The datasets used in this study have been progressively transformed to do more detailed mining process. Although the details will be presented in next section, we can advance that we have used 3 datasets derived from the original one.

The first dataset, with 116 examples (students), is described by 3 attributes (marks achieved in every test) and a binary nominal class (passing the subject or failing it – including absent students –). In Fig. 1 it is shown the class distribution for three different marks.

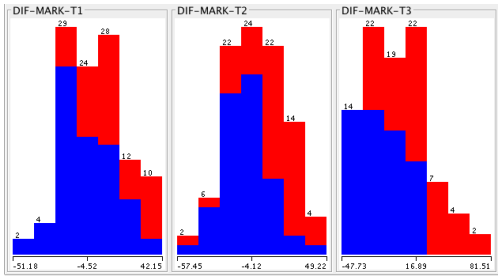


Fig. 2. Class distribution visualizing marks and differences with the average marks (for tests T1, T2 and T3) in the second dataset (116 students). Blue color represents students that fail the evaluation (or absent themselves) and red color represents students that pass the evaluation. This chart is plotted by the Weka framework.

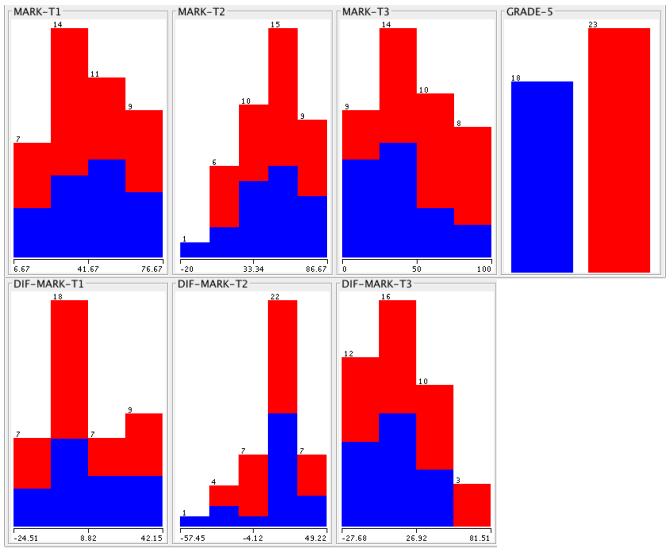


Fig. 3. Class distribution visualizing marks and differences with the average marks (for tests T1, T2 and T3) in the third dataset (41 students). Blue color represents students that have grades D or E and red color represents students that have grades A, B or C. This chart is plotted by the Weka framework.

In the next step we calculated the differences between the mark itself and the average valued achieved in that test by students during their course. Therefore, we incorporated 3 new attributes to the dataset. In Fig. 2 it is shown the class distribution for such new attributes.

Finally, once we have detected patterns to separate students that pass the evaluation and those students that do not pass it, we were interested in inducing some models that could find some pattern to differentiate between best students (with A, B or C grades) and the rest of students that pass the evaluation (D or E grades). In this dataset we only had 41 students so the induction algorithm had some problems with so few examples. To solve it we resample the dataset [14] making it five times bigger (205 examples) and configured J48 to examine a bigger number of examples before expanding (minimum of 20 examples) in order to avoid overfitting and reduce the complexity of the model [12]. In Fig. 3 it is shown the class distribution for this last dataset.

III. RESULTS

In this section we present the results that we have collected

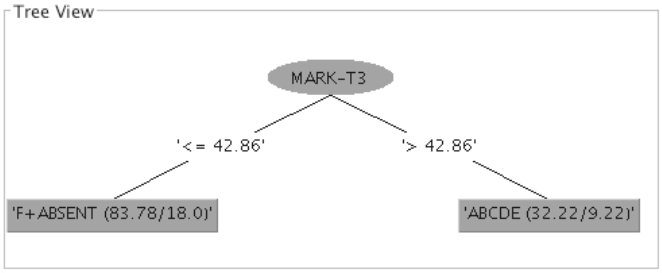


Fig. 4. Decision tree induced by J48 using the first dataset. Attributes are the marks for T1, T2 and T3; the binary class separate between students that pass (ABCDE) or not pass (F+ABSENT) the global evaluation process.

after applying the mining process to the data previously described. As we have explained, we have used a TDIDT algorithm (J48 implementation of C4.5), so the induced models are decision trees, what make possible an easy interpretation of the patterns. In addition, we can rely on the results, because validation processes show high confidence levels. The validation processes we have conducted are 10-fold cross validations.

For the first dataset, that which separates students in a binary class (pass or not pass the evaluation) and only include the marks achieved for every test (T1, T2 and T3), the pattern is easy to understand (even no TDIDT algorithm would be necessary because the class distribution in Fig. 1 shows a similar information). The most important attribute to determine the difference between two student profiles is the mark achieved for the last test (T3), the most close to the final exam. The decision tree, shown in Fig. 4, is not surprising, but reflects the ability of machine learning algorithms to find patterns. Furthermore, the validation shows 80% accuracy, quite reliable considering the number of examples and the class unbalance.

Analyzing the second dataset, extended with new attributes that summarize the differences between the own mark and the average value, some additional knowledge is extracted. Decision tree (Fig. 5) reveals that once we know the mark for T3 (root node), we can detect some other differences. In this case, the new added attributes reveal as important elements to determine the final achievement of the students. Particularly students that are below 42.86 points in T3, need to do best that the average in T1 and T2 to pass. So the requirements are not

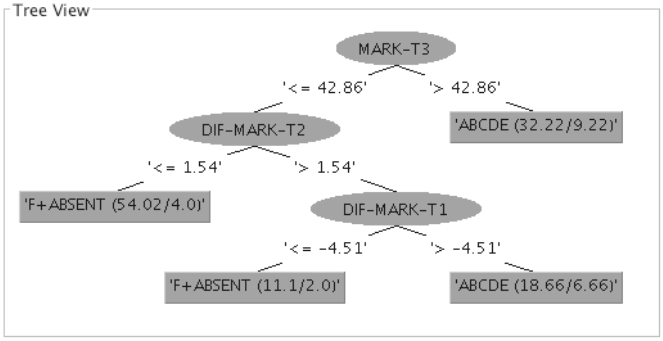


Fig. 5. Decision tree induced by J48 using the second dataset. Attributes are the marks (for T1, T2 and T3) and the difference with the average value; the binary class separate between students that pass (ABCDE) or not pass (F+ABSENT) the global evaluation process.

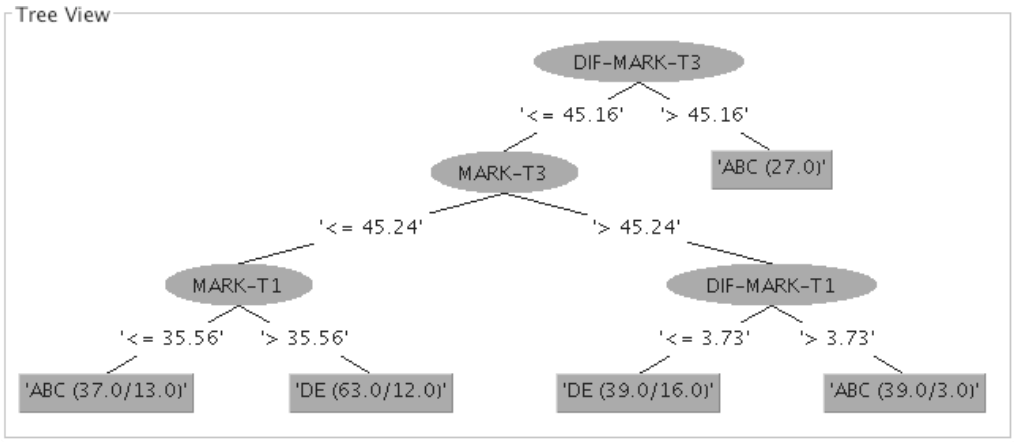


Fig. 6. Decision tree induced by J48 using the third dataset. Attributes are the marks (for T1, T2 and T3) and the difference with the average value; the binary class separate between students that achieve best grades (ABC) and the other ones (DE).

so restrictive for them, note that students do not need to pass the test (50 points out of 100), they only need to do best that the average value (close to 35 out of 100, see Table I)

Once again, the validation presents a quite reliable model (even higher than previous model) because we have 87% accuracy. This makes sense because we have added new attributes that help to better differentiate between student profiles.

Finally, once we have identified some criteria that determine differences between students that pass or not pass the final evaluation, we focus in those that pass the evaluation and how good their results are. Concretely we want to know if there is some element that reveals how they differ. In Fig. 6 we show the decision tree induced by J48 which reliability is relatively high (close to 80% accuracy).

Once again, the last test (T3) seems the most decisive element. It is logical, because this test includes and extends the concepts and abilities needed for the second test (T2). But this time the model differs substantially from previous ones because the actually important attribute is not the mark itself, but the difference with the average value. For every student (there is no exception, see most right-side branch in the decision tree) which T3's mark is beyond the average value in more than 45.16 points (out of 100), the final grade is better than D (A, B or C grade). Note that this difference is even greater than the standard deviation (27.84).

For those students that do not surpass the average value in such quantity, we find both kinds of students. In this case, differences between them are less clear and they could be even misunderstanding at a first moment. As it can be seen in the decision tree, first test information (T1) is selected to expand the tree in the deepest levels. It seems strange that students with lower marks ($\leq$ 35.56) get highest grades in the final evaluation, but we found some explanations that diminish the importance of such strangeness. On one hand, we can see that such asseveration is not so strong, because not all the examples are correctly classified (see <second> number in leaves), so some level of noise is present in that attribute. On the other hand, if we know that first test (T1) is conducted at the beginning of the semester and its relation to final exam is very poor, we can think that dependencies are arguable; even more, we can suppose that good students with a "poor" mark in the first test can detect the necessity of strengthen the efforts because they did an incorrect initial calibration about the difficulty of the subject.

IV. CONCLUSION

In this paper we have studied, by using data mining techniques, the possibility that learning processes in the academic context could incorporate new and relevant knowledge that enables improvements in such processes.

In the conducted analysis we have detected that there are relations between the continual assessment carried out during the semester and the final evaluation. These relations, correctly used, can lead the adaptation of existing strategies or to boost the integration of new methods in subjects for future courses.

To a large extent, such improvements depend on having enough data about the evolution of the evaluations, on analyzing them continuously, on detecting anomalous behaviors; and on developing preventive and corrective actions (new exercises, individual tutorial actions, etc.). At this moment, Web-based Educational Systems offer tools to obtain and process that data, so its usage is highly recommended.

In addition, due to the flexibility of these systems, they can be adapted and extended. New functionalities can be added, and two different developments can be incorporated to progress in the previously mentioned improvements. As a first point, Web-based Educational Systems can collect more data, those that have shown their usefulness for data mining analysis (even calculating new fields). As a second feature, they could incorporate the mining process in the core of the system in order to offer a dual advantage: helping the teacher with the analysis tasks (assessment task) and helping the students by guiding their learning process (adapting task).

This study reveals many future research lines in different dimensions. There exists a wide diversity of techniques in the data mining field, so selecting other paradigms could improve the knowledge acquired (association rules, decision rules,

etc.). If we are not so interested in the understandable knowledge (assessment task) and we prefer to provide the system with better guiding characteristics (adapting task), we have a perspective even broader because we could use many other strategies not so easily human-readable but very accurate (ensembles, neural networks, etc.).

Another promising area is the automatic or semi-automatic tune up of the Web-based Educational Systems. It is interesting to modify the educational system to respond to specific necessities of students [15]. This adaptation could even be implemented in real time, responding during the interaction with the student.

In this sense there are emerging new areas in machine learning and data mining related with data streams [16], very large (even non-ended) datasets that grow increasingly. Its usage fits very well with the dynamic of Web-based Educational Systems that are open constantly and can interact with students (and receive data) at every moment. Therefore, incorporating incremental algorithms [17] that can learn in this context would be positive. Additionally, as the student profile is not static, providing mechanisms to detect concept drift [18, 19] would contribute to create much more adaptable systems.

REFERENCES

[1] P. Brusilovsky, E. Schwarz and G. Weber, "ELM-ART: An intelligent tutoring system on World Wide Web", *Lecture Notes in Computer Science*, vol. 1086, pp. 261-269. 1996.

[2] C. Romero, S. Ventura, A. Zafra and P. de Bra, "Applying Web usage mining for personalizing hyperlinks in Web-based adaptive educational systems", *Computers & Education*, vol. 53, pp. 828–840. 2009.

[3] E. Melis, E. Andrès, J. Büdenbender, A. Frishauf, G. Goguadse, P. Libbrecht, M. Pollet and C. Ullrich, "ActiveMath: A generic and adaptive web-based learning environment", *International Journal of Artificial Intelligence in Education*, vol. 12, pp. 385-407. 2001.

[4] P. Brusilovsky and C. Peylo, "Adaptive and Intelligent Web-based Educational Systems", International Journal of Artificial Intelligence in Education, vol. 13, pp. 156-169. 2003.

[5] U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, "From Data Mining to Knowledge Discovery in Databases", AI Magazine, vol. 17, pp. 37-54. 1996.

[6] C. Romero and S. Ventura, "Educational data mining: a review of the state of the art", *IEEE Transactions on Systems, Man, and Cybernetics-Part C*, vol. 40, pp. 601-618. 2010.

[7] E. Guzmán, R. Conejo, and J. Gálvez, "A Data-Driven Technique for Misconception Elicitation", *Lecture Notes in Computer Science*, vol. 6075, pp. 243-254, 2010.

[8] C. Romero, S. Ventura, and P. De Bra, "Knowledge discovery with genetic programming for providing feedback to courseware author", *User Model. User-Adapted Interaction*, vol. 14, pp. 425-464. 2004.

[9] R. Conejo, E. Guzmán, E. Millán, M. Trella, J. L. Pérez-De-La-Cruz and A. Ríos, "SIETTE: A Web-Based Tool for Adaptive Testing", *International Journal of Artificial Intelligence in Education*, vol. 14, pp. 29-61. 2004.

[10] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, I. H. Witten, "The WEKA Data Mining Software: An Update", *SIGKDD Explorations*, vol. 11, pp. 10-18. 2009.

[11] European Communities, "*ECTS Users' Guide*", Luxembourg: Office for Official Publications of the European Communities, 2009.

[12] T. M. Mitchell, *Machine Learning*. New York, McGraw-Hill, 1997, ch. 3.

[13] J. R. Quinlan, *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA. 1993.

[14] P. H. Lee, "Resampling Methods Improve the Predictive Power of Modeling in Class-Imbalanced Datasets", *International Journal of Environmental Research and Public Health*, vol. 11, pp. 9776–9789. 2014

[15] A. Kozierkiewicz-Hetmańska and N. Nguyen, "A method for learning scenario determination and modification in intelligent tutoring systems", *International Journal of Applied Mathematics and Computer Science*, vol. 21, pp. 69-82. 2011.

[16] J. Gama, *Knowledge Discovery from Data Streams*. Chapman & Hall/CRC. 2010.

[17] J. del Campo-Ávila, G. Ramos-Jiménez, J. Gama and R. Morales-Bueno, "Improving the performance of an incremental algorithm driven by error margins", *Intelligent Data Analysis*, vol. 12, pp. 305-318. 2008.

[18] J. Gama, I. Žliobaitė, A. Bifet, M. Pechenizkiy and A. Bouchachia, "A survey on concept drift adaptation", *ACM Computing Surveys*, vol. 46 (4), Article 44, 37 pages. 2014.

[19] I. Frías-Blanco, J. del Campo-Ávila, G. Ramos-Jiménez, R. Morales-Bueno, A. Ortíz-Díaz and Y. Caballero-Mota, "Online and non-parametric drift detection methods based on Hoeffding's bounds", *IEEE Transactions on Knowledge & Data Engineering*, to be published.

José del Campo-Ávila received the Ph.D. degree in Software Engineering and Artificial Intelligence from the University of Málaga in 2007. He is an Associated Professor in the Languages and Computer Science Department of the University of Málaga, Málaga, Spain, where he has worked since 2003. His research interests include incremental learning, mining data streams for classification, and multiple classifier systems, among others. Dr. del Campo-Ávila is member of the Research and Applications in Artificial Intelligence Group. He is also a regular member of program committees of international conferences, such as Intelligent Data Analysis (IDA) or Data Stream track in ACM Symposium on Applied Computing (DS-SAC).

Ricardo Conejo received the Ph.D. degree in Ingeniero de Caminos, Canales y Puertos (Civil engineer) from the Technical University of Madrid, in 1995. He is a Full Professor in the Languages and Computer Science Department of the University of Málaga, Málaga, Spain, where he has worked since 1986. In addition, he is the Director of the Research and Applications in Artificial Intelligence Group at the same university. He has 25 years of experience in the development of intelligent tutoring systems. He has published extensively on this field. His research interests currently focus on adaptive testing, student knowledge diagnosis and intelligent tutoring systems. He has also worked on fuzzy logic, model-based diagnosis, multiagent systems, and artificial intelligence applied to civil engineering.

Prof. Conejo is a regular member of program committees of international conferences, such as User Modeling Adaptation, and Personalization (UMAP), Intelligent Tutoring Systems (ITS), and Artificial Intelligence in Education (AIED).

Francisco Triguero was born in Málaga, Spain, in 1955. He received the Mathematics Degree, in 1977 from the University of Málaga, Spain, and the Ph.D. degree, in 1982 from the University of Granada, Spain. He works in intelligent tutoring systems, knowledge discovery, and computational learning.

He was a Lecturer in the Mathematics Department of Málaga University (1977–1980), Assistant Professor in the Mathematics Department of Málaga University (1980–1984), Full Professor of Business Mathematics at the University of Seville (1984–1986), and Full Professor of Business Mathematics at the University of Málaga since 1986. He has occupied the position of Director of the Computer Service of Málaga University. Founder and first Director of the Computer Science School and the Telecommunication School, in 1987, first elected Director of the Computer Science School, in 2000, President Chair of the CAEPIA'97 (Conference of AEPIA), and Cochairman of ICALP'2002 and AH'2002.

Dr. Triguero is a member of EATCS (European Association for Theoretical Computer Science) and AEPIA (Spanish Association for Artificial Intelligence), member of the Andalusian Universities Council, and Member of the Advice Council of the Andalusian Institute of Development.

Rafael Morales-Bueno was born in Málaga, Spain, in 1955. He received the Mathematics Degree, in 1978 from the University Complutense of Madrid, Spain, and the Ph.D. degree, in 1991 from the University of Malaga, Spain. He works in data mining, knowledge discovery, computational learning and also in computability and complexity (from classical and fuzzy point of view), among others. He was an Assistant Professor in Computer Science Department of Málaga University (1984–1987), Associate Professor (1987-2009), and Full Professor of Computer Science (2009-). He has occupied the position of Director of the Computer Science School (2000-2004) and Vice-rector of Infrastructures (2004-2012). Chairman of ICALP'2002 and other conferences. Dr. Morales-Bueno is a member of the Research and Applications in Artificial Intelligence Group. He is also a Member of EATCS (European Association for Theoretical Computer Science) and AEPIA (Spanish Association for Artificial Intelligence). Rafael Morales-Bueno is also Vice-president of Momopocket EDE S.L. (Electronic money Entity), a company specialized in mobile payment systems.

A Repository of Semantic Open EHR Archetypes

Fernando Sánchez, Samuel Benavides, Fernando Moreno, Guillermo Garzón, Maria del Mar Roldan-Garcia, Ismael Navas-Delgado, Jose F. Aldana-Montes

University of Málaga, Malaga, Spain

***Abstract* — This paper describes a repository of openEHR archetypes that have been translated to OWL. In the work presented here, five different CKMs (Clinical Knowledge Managers) have been downloaded and the archetypes have been translated to OWL. This translation is based on an existing translator that has been improved to solve programming problems with certain structures. As part of the repository a tool has been developed to keep it always up-to-date. So, any change in one of the CKMs (addition, elimination or even change of an archetype) will involve translating the changed archetypes once more. The repository is accessible through a Web interface (http://www.openehr.es/).**

***Keywords* — openEHR, Semantic Web, CKM**

I. Introduction

As stated in [1] EHRs (Electronic Health Records) and ePrescribing have a real impact in the healthcare at a service level and also at the economical level. However the economic impact is reflected in the net benefits only in an average period of 7 years. The use of standards for the establishment of EHRs in healthcare systems would reduce this latency period.

The development of health information systems has been guided by the need for health systems to manage the huge amounts of information that make the use of physical methods unfeasible. However, these systems are not usually constrained to standards. Thus, different hospitals working together or even different services within the same hospital cannot share information about their patients.

Most advanced EHR architectures and standards are based on the dual model-based architecture, which defines two conceptual levels [2].

OpenEHR has at its core the aim of providing the necessary elements for managing electronic health records, providing ways of modelling all the agents implied in a health environment. The openEHR Foundation provides specifications which define a health information reference model together with a language for developing archetypes (clinical models). This language is not part of any software or query language by default. This architecture, based on archetypes, enables the use of external health terminologies (SNOMED CT, LOINC and ICD). OpenEHR uses the dual-model architecture, which has also influenced HL7 CDA. In dual model approaches, archetypes constitute a tool for building clinical consensus and this enables interoperability between different health information systems.

In this approach we are working towards extending how the models are published by providing new perspectives in the use of OWL as a language to provide semantically rich clinical models. Using a translator, we have built a repository of OWL models derived from public ADL models. Ongoing work is helping this proposal to provide ways of improving this semantics by aligning archetypes and health records with ICD-10 and SNOMED-CT. However, because the structure of the EHR is annotated with such terminologies, the information contained in an EHR is mostly composed of text descriptions without terminology annotations on the patient data.

Section 2 presents some related work. Section 3 describes the archetype translation process. Section 4 presents the current version of the repository and its user interface, to conclude with Section 5 explaining the main conclusions and ongoing work.

II. Related Work

Archetypes are considered an important element in the achievement of the semantic interoperability between EHR systems. So, the design of methods to manage them is fundamental [3]. The translation of openEHR archetypes to OWL is not a novel proposal. [4] presents the first proposal of an ontology for representing archetypes in OWL. This ontology is divided into seven integrated ontologies:

- EHR EXTRACT Reference Model. It defines the semantics shared by all kinds of Extract requests and Extracts from openEHR data.
- EHR Reference Model. It contains a representation in OWL of the information model of the openEHR EHR.
- Data Structures Reference Model. It represents the shared data structures used in openEHR reference model, including lists, tables, trees, and history, together with one possible data representation (hierarchical).

DOI: 10.9781/ijimai.2015.327

DOI: 10.9781/ijimai.2015.326

- Support Reference Model. It defines identifiers, assumed types, and terminology interface specification used by openEHR reference model.
- Common Reference Model. It contains shared concepts, including the archetype-enabling LOCATABLE class, party references, audits and attestations, change control, and authored resources.
- Demographic Reference Model. It describes the architecture of the openEHR Demographic Information
- Data Types Reference Model. It represents data types, including quantities, date/times, plain and coded text, time specification, multimedia and URIs.

Figure 1 shows a part of this ontology. As can be observed, the design of this ontology is directly driven by the syntactic structure of the archetypes, including their main types, without taking into account compressibility or reusability. From a semantics point of view, this is an inconsistent ontology (tested using the Pellet reasoner in Protégé 4.3), so it cannot be used for reasoning purposes. However, the positive aspect of this ontology is that it is complemented by translation software [5] for obtaining OWL versions of ADL archetypes. This translator is based on the ADL API and the Archetype Object Model (AOM). The OWL model is built using Jena to construct the ontology model in memory while the ADL archetype is simultaneously parsed. A negative aspect of this translator is that it only includes the translation of 2 of the 4 archetype types, and many of the archetypes in these two types cannot be translated due to programming errors.

In this paper we present the roadmap from this approach to reach some goals:

- A comprehensible and reusable consistent OWL ontology.
- A complete translator for any ADL archetypes to consistent OWL ontologies.
- A repository of archetypes and translations able to trace the evolution of the archetypes.
- Software able to automatically align clinical records with external vocabularies.

III. Archetype Translation

An archetype constrains the entities of the reference model. The constraints are applied to the attributes defined for each entity: range, cardinality, etc. In this way, each constrained entity is defined by means of an OWL class in which the corresponding constraints are defined [6].Using the existing translator we have taken several steps to improve it.

A. Error detection.

In this step we have tested the translator using public archetypes in the openEHR CKM (http://www.openehr.org/ckm/). The automatic execution of these archetypes showed the following errors that were solved on the translator provided in our portal (http://www.openehr.es/):

- Non-existing nodes. Some ADL nodes were not expected at certain parsing steps, and this lead the software to an error, stopping the translation process. These nodes were analysed and the translator extended to deal with them properly.
- Repeated class names. The names of the classes in the translation directly rep-resent ADL nodes. ADL does not prevent us from using the same name for different nodes, but OWL does not allow the use of the same name in different classes. In order to solve this problem, the names for these classes were automatically detected and changed to a new name using the parent class name as a prefix.

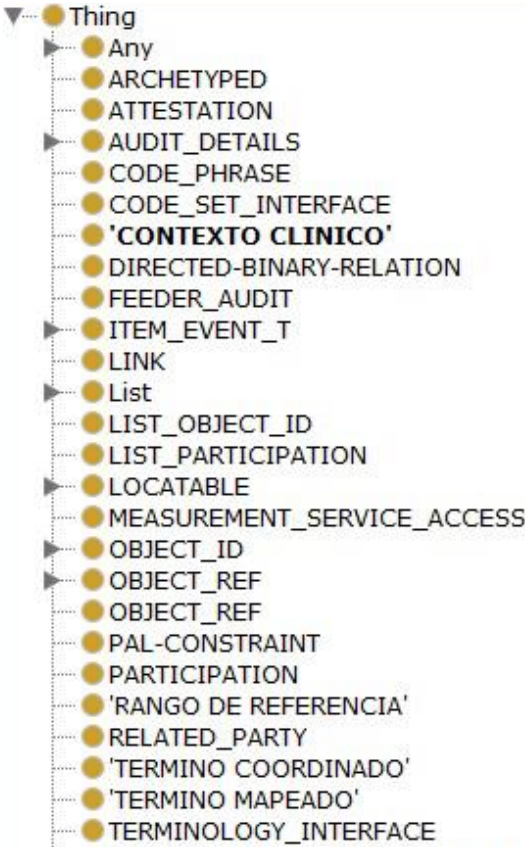


Fig. 1. Main structure of the archetype ontology

B. Incompleteness.

The translator has been shown to be incomplete and so unable to deal with over half of the public archetypes we wished to translate automatically.

- Archetype types that are not translated. The original translator does not trans-late ACTION AND INSTRUCTION archetypes. INSTRUCTION and ACTION have been added to the reference ontology and now the archetypes in this cate-gory are properly translated.
- External vocabulary annotations (SNOMED, ICD, etc.) are not translated. The first step in solving this issue has been to add a new concept to the resulting ontology:

ONTOLOGY_CONCEPT. Thus, the translator has been extended with a component for detecting and dealing with external vocabulary annotations. When the ADL parser detects these annotations this component is activated to add a new instance of the new ONTOLOGY_CONCEPT indicating the external vocabulary used (SNOMED, ICD, etc.) and the term is referenced in the ADL annotation. These annotations will be of help when trying to align clinical data with external vocabularies as this will provide a context to be used by the text mining process.

C. Improve the resulting OWL ontology.

The translator is being modified to eliminate the generation of unnecessary nodes. Some of the concepts added to the OWL ontology were direct translations from the ADL language and are not needed to rep-resent the information of the archetypes. This part of the translator is being modified to use a different structure of the OWL ontology without using these intermediate class names, reducing the complexity of the resulting OWL ontology. This modification which will lead to a totally different translator is still ongoing work which will describe in the following sections.

D. Test case generation.

The translation of archetypes to OWL enables the possibility of using RDF Database Management Systems to deal with clinical data represented as instances (individuals) of these OWL ontologies. However, there are no examples of how clinical data should be represented in these ontologies. Thus, we have developed an instance generator to provide test cases for the data management. Our instance generator asserts individuals in a given ontology in two different ways: inserting individuals according to certain data or inserting individuals randomly generated in a given range.

In order to insert individuals by given data we should follow these steps:

- Instance the reference ontology using "columnX" where "X" is the number of the column from which the program should take the data. We should keep in mind that the first column is "column0".
- The name of each instance in the reference ontology has to be given, with its version at the end, e.g. "example.1", "example.2", "example.1.1", "example.2.3.4", [...].
- This algorithm can be configured to take input files, and decide where to write the results. The separator of data by default is tabulator.
- The program will insert as many individuals as there are lines in the input file.

In order to insert randomly generated individuals in a given range we should follow these steps:

- The input file must have a first line with the type of value that we would like to use separated by spaces (being I=Integer, D=Double and S=String.
- The input file must include one line for each of the types we put in the first line.
- The reference ontology is instanced in the same way as in the previous case, but the data is not collected from the input file, rather it is randomly generated, taking the data types indicated.
- The output is a file that can be used as input for the previous case. Thus, it is possible to create a workflow that uses both cases together, although their maintenance is independent. So, the changes in the reference ontology will only affect the first case, but not the second.

E. Translation examples

Current translation implies that the result is an ontology with a similar structure as an ADL file. Thus, a simple ADL file (Figure 2) will produce a complex structure based on subsumption and object properties.

Fig. 2. Autopsy observation archetype

The generation of instances will produce instances for the whole structure of the given archetype ontology. This means generating a lot of instances for intermediate concepts that will serve only as the connection between the archetype and the given data. For example, the following input file (for Blood Pressure archetype) will generate a complex structure of instances as shown in Figure 3:

```
13.10      0 mm[hg] 8.100 mm[hg]  9.50  0 mm[hg] 75

13.20      0 mm[hg] 8.200 mm[hg]  9.60  0 mm[hg] 85

13.30      0 mm[hg] 8.300 mm[hg]  9.70  0 mm[hg] 95

13.40      0 mm[hg] 8.400 mm[hg]  9.80  0 mm[hg] 105

13.50      0 mm[hg] 8.500 mm[hg]  9.90  0 mm[hg] 115

13.60      0 mm[hg] 8.600 mm[hg]  9.00  0 mm[hg] 125
```

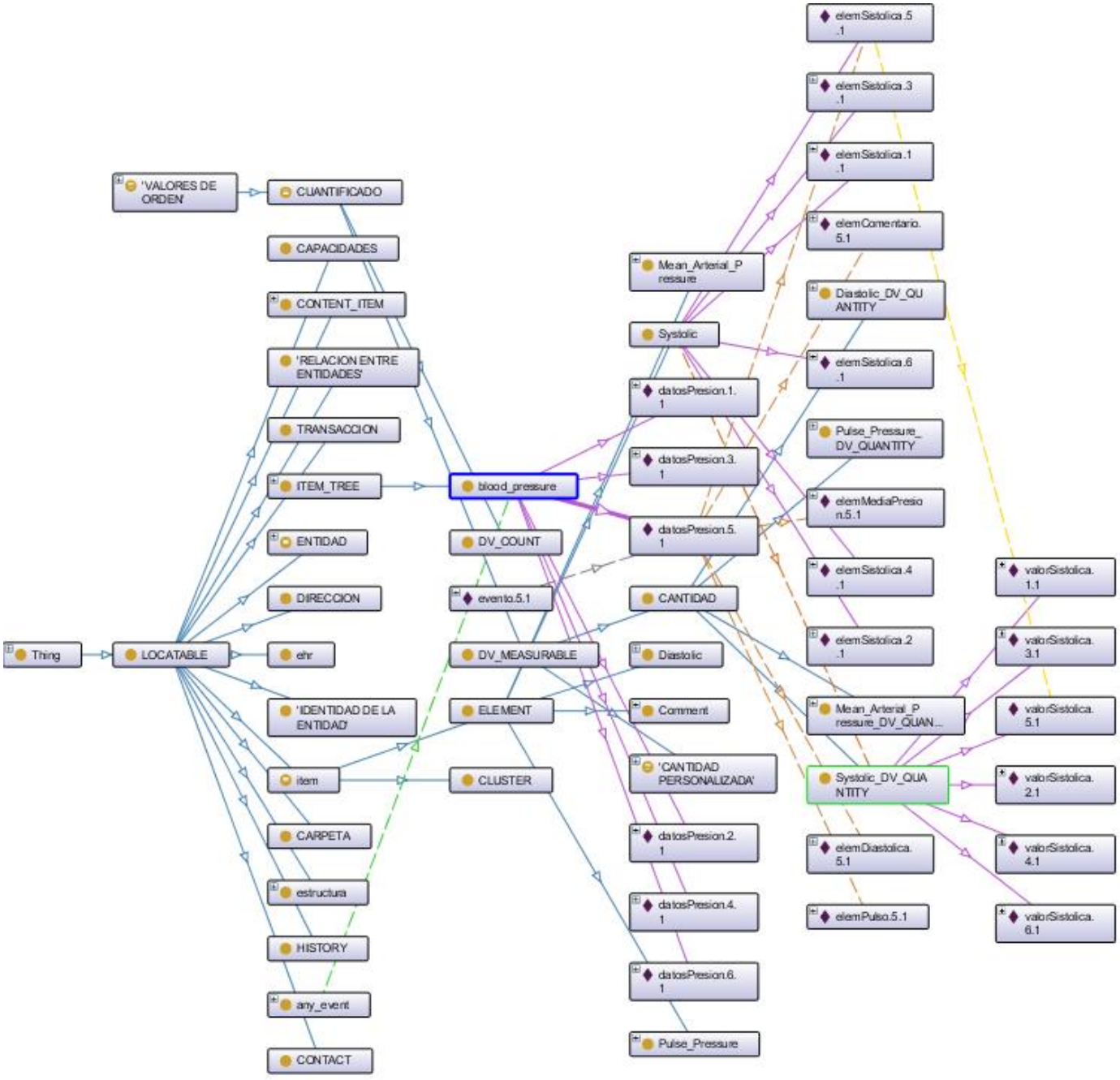


Fig. 1. Example of instances for a given data file

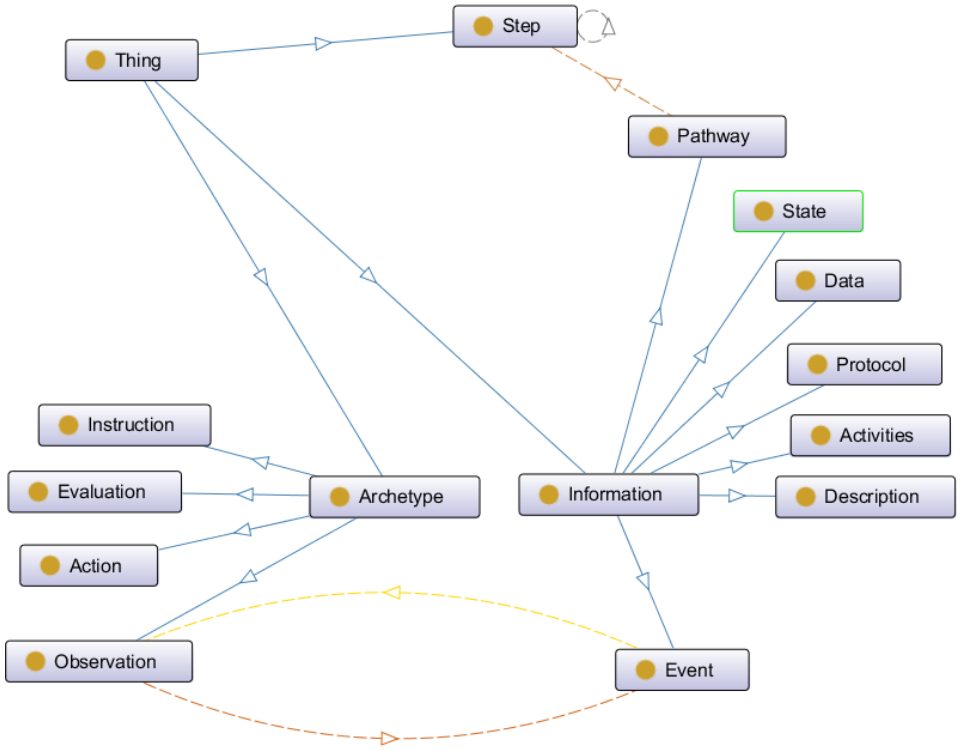


Fig. 2. Reference OWL ontology

F. 3.2 Translation results

An archetype is consistent if its set of defined constraints over both the reference model and the parent archetype are satisfiable. It is necessary to analyse the results of the translation and to check the quality of the archetypes represented in OWL. Generated instances for the current archetypes have been manually evaluated to discover translation errors. This manual process is based on the comparison of the translated archetype as an OWL ontology with the original version in ADL. Nodes are compared by their name and relationships with the other nodes. This ensures that although at first glance the archetype represented in OWL seems to have been translated correctly, there are no hidden translation failures. The quality of the translation is an important part of the translation process in order to ensure a certain level of quality of the translations offered.

IV. Repository Management

The repository of archetypes is built and updated using a daily batch process connecting to a list of CKMs. This process checks all the archetypes contained in the external repositories, extracts them and compares the contents of the CKM with the local repository. If there are any differences, the process updates the archetypes in the local repository and translates the modified ones to OWL.

In order to connect to the CKM the system uses a web service that provides the CKM and returns a compressed file with all archetypes structured in directories, classified by type. The following CKMs are currently being accessed:

- NEHTA = http: //dcm.nehta.org.au/ckm/
- openEHR = http: //www.openehr.org/ckm/
- uk = http: //clinicalmodels.org.uk/ckm/
- ezdrav = http: //ukz.ezdrav.si/ckm/
- russia = http: //simickm.ru/ckm/

An archetype can pass through several states (initial, draft, review team, etc.). If an archetype is “published”, it cannot be modified. In this case, modifications should be done as an archetype with the same name and higher version number. This way of managing CKM prevents the modification of published archetype contents. The contribution of updating the repository is to keep all versions of archetypes to provide users with translations to the archetype version they are using in their Health Information System, even if a new version has been published. The synchronisation process is as follows:

- If a file has been modified internally, it is replaced in the local repository by the new one and the conversion to OWL is deleted.
- If a new archetype appears, then it is copied to the local repository, this occurs when a new archetype is created in the CKM or is versioned.
- Archetypes are not deleted from the CKM rather they are labeled as rejected or obsolete. Thus, it is not necessary to check whether an archetype is missing from the local repositories.

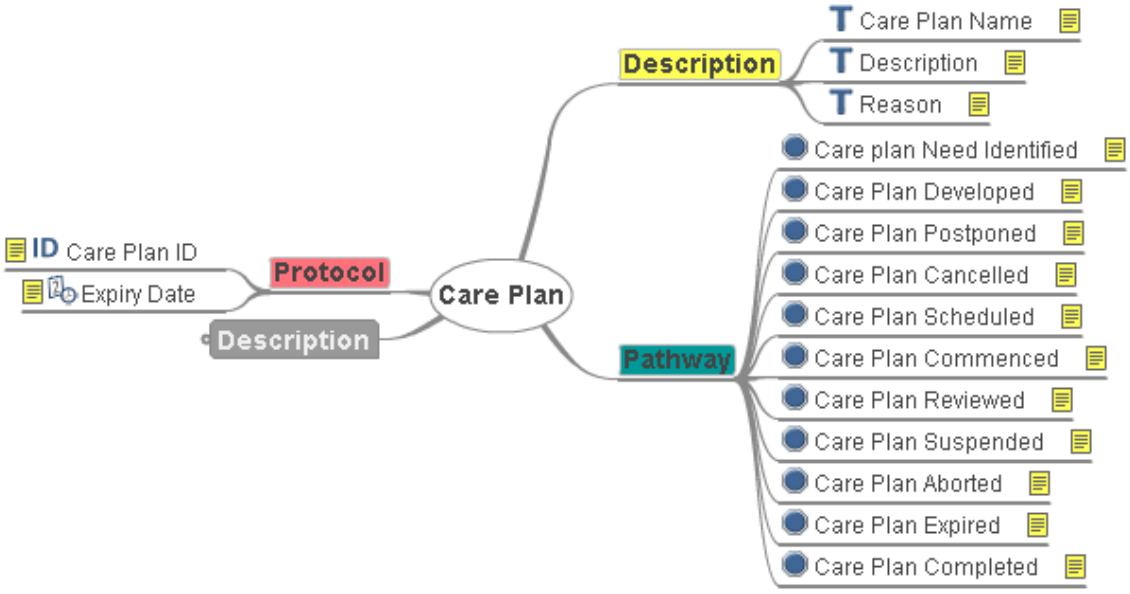


Fig. 3. Care Plan archetype

- Once the local and external repositories have been synchronised, the OWL translation process checks for each of the archetypes added or modified in the local repository.

The repository contains the automatically translated archetypes from public archetype repositories like, for example, CKM. However, ADL allows users to define their own archetypes. For this reason, the translation tool is included in the portal, so users can test its functionality. The system does not keep a copy of the archetype, or the translation, unless the user asks for them to be included in our repository.

V. Discussion and Conclusions

The use of standards such as OpenEHR will reduce the time to return on the investment of putting an EHR system to work, with the corresponding economic impact. Additionally, the use of semantics opens new ways of interoperability even with other standards making worthy this initial economic effort. The automatic translation of openEHR archetypes to OWL has been approached in the past. However, in the cur-rent climate in which the interoperability of health information systems is a priority, this topic is of strong interest. For this reason, we have started with previous work and analysed the existing problems in these types of translations. Some of the problems detected have been solved, and an improved version of the translator has been used to provide a repository of OWL ontologies representing public archetypes.

However, there is still much work to be done in this approach. The main issue we are addressing is the design of a reference OWL ontology to lead the translation process towards consistent, comprehensible and reusable ontologies. The reference mod-el (Figure 4) we have designed in the first phase simplifies the representation of archetypes. For example, for the Care Plan Archetype (Figure 5), the translation would be similar to the ontology in Figure 6.

Neither of the formal representations of ICD-10 presented in the literature has been classified nor their consistencies checked. Even more, some of then uses an OWL-Full component that prevents its use in a semantic classification system based on reasoning. Other approaches propose to model the ICD-10 exclusions using the owl:disjoint axiom, that could lead to a loss of important information and generate inconsistences in the model. There are no ontologies that combine SNOMED-CT and ICD-10-CM. SNOMED-CT and ICD-10 are broadly used in the field of medicine. In fact, SNOMED-CT is being used in most of the Health Information Systems. For this reason, our research group is working on modelling the ICD-10 (International Classification of Diseases, 10th version) [7] as an OWL ontology [8]. This medical classification standard, maintained and published by the WHO (World Health Organization) is used to classify diseases and health problems that have been recorded on death certificates and also in other records. Our ontology has also been aligned with SNOMED-CT [9]. SNOMED-CT terminology often referenced as an ontology, includes all those concepts that relate to each other logically within a specific domain [10]. As many openEHR archetypes are annotated with an ICD-10 code, this enables the possibility of aligning the OWL ontologies in our repository with our ICD-10 ontology. By means of this alignment, the reasoning capabilities of the OWL language can be exploited so as to obtain implicit information about the clinical concept described by the archetype, based on the information contained in ICD-10 and SNOMED-CT, such as its relationships with other clinical concepts, diseases and clinical procedures, to name a few.

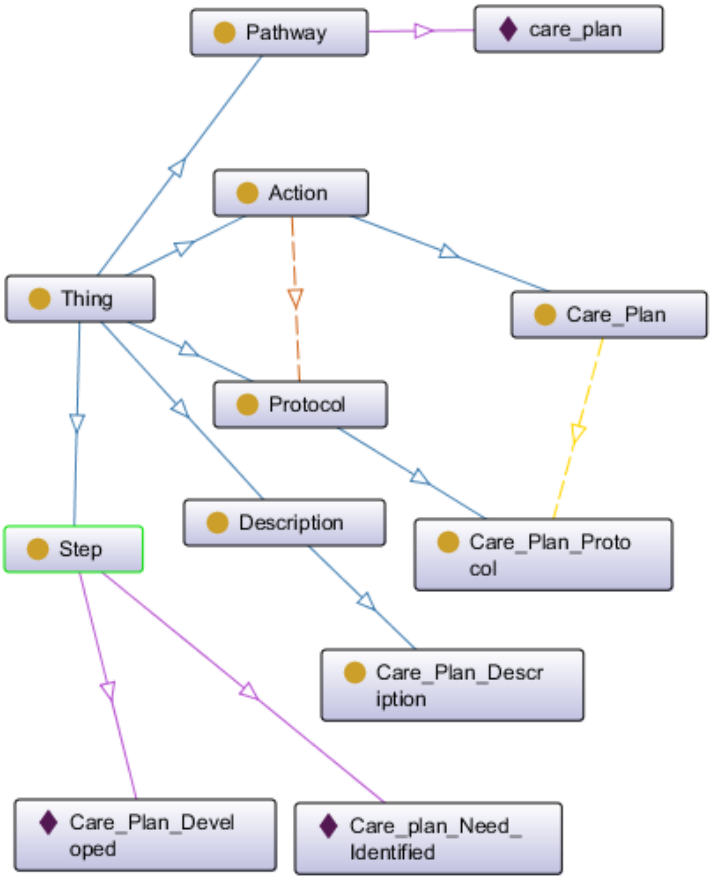


Fig. 4. Translation of Care Plan archetype

REFERENCES

[1] The socio-economic impact of interoperable electronic health record (EHR) and ePrescribing systems in Europe and beyond. http://www.ehr-impact.eu/downloads/documents/EHRI_final_report_2009.pdf

[2] Catalina Martinez-Costa; Marcos Menárguez-Tortosa; Jesualdo Tomás Fernández-Breis. An approach for the semantic interoperability of ISO EN 13606 and OpenEHR archetypes. Journal of Biomedical Informatics. October 2010.

[3] María del Carmen Legaz-García; Catalina Martinez-Costa; Marcos Menárguez-Tortosa; Jesualdo Tomás Fernández-Breis. Exploitation of ontologies for the management of clinical archetypes in ArchMS. University of Murcia (España), 2012.

[4] Isabel Román Martínez. Aportaciones metodológicas a la integración de sistemas de informa-ción sanitarios basada en gestión semántica. Supervisors: Rafael Robles Arias, Germán Madina-beitia Luque. University of Sevilla (España), 2006.

[5] Lezcano Matías, Leonardo. Combining ontologies and rules with clinical archetypes. Universi-dad de Alcalá. Departamento de Ciencias de la Computación. Supervisor: Sicilia Urbán, Miguel Ángel. 2012.

[6] Marcos Menárquez-Tortosa; Jesualdo Tomás Fernández-Breis. OWL-based reasoning methos for validating archetypes. Journal of Biomedical Informatics. November 2012.

[7] International Classication of Diseases (ICD) website, 775 http://www.who.int/classifications/icd/en/.

[8] Pastor-Rubia, D.; Rold´an-Garca, M.M.; Navas-Delgado, I. Aldana-Montes, J. F. DTEKT: semantic classification for medical diagnosis Riccardo Bellazzi and Paolo Romano (eds.). Clinical bioinformatics. 11th international workshop. NETTAB 2011. Network tools and application in biology. October 12th-14th 2011, Collegio Ghislieri, Pavia, Italy, 2011, 86-87.

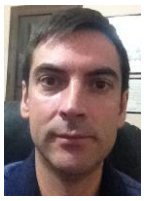
[9] SNOMED-CT website, http://www.ihtsdo.org/snomed-ct/.

[10] Jose Luis Allones; David Penas; María Taboada; Diego Martinez; Serafín Tellado. University of Santiago de Compostela (España). G. O. Young, "Synthetic structure of industrial plastics (Book style with paper title and editor)," in *Plastics*, 2nd ed. vol. 3, J. Peters, Ed. New York: McGraw-Hill, 1964, pp. 15–64.

José F. Aldana-Montes is Full Professor at the University of Málaga, Spain. He is the main researcher of the Khaos Research Group. His research focuses on management, integration and analysis of data.

Ismael Navas-Delgado is PhD and Assistant Professor at the University of Málaga, Spain. His research focuses on the use of Semantics and Big Data technologies in Life Sciences.

Maria del Mar Roldan-Garcia had her PhD in Computer Science in 2012 by the University of Málaga, Spain. Currently she is a Lecturer at the same University. Her main interest are ontologies and Semantic Web, scalable semantic reasoning, smart data and biomedical ontologies development and applications.



Fernando Sánchez is Technical Engineer in Computer (Universidad de Málaga, 1997) and MSc in Software Engineering & Artificial intelligence (Universidad de Málaga, 2014). His research is focused on electronic health records and openEHR standard.

Samuel Benavides is student of Bioinformatics at University of Málaga. His major interests are text mining and semantic web. Currently, he is a grant holder working at Khaos, a research group by the University of Málaga.

Fernando Moreno is student of Bioinformatics at University of Málaga. His major interests are data mining, algorithmic and metaheuristics. Currently, he is a grant holder working at Khaos, a research group by the University of Málaga.

Guillermo Garzón Sánchez-Cruzado is studying the degree of Health Engineering the specialization of Bioinformatics Engineering by University of Málaga and University Sevilla. Currently, he is a grant holder working at Khaos, a research group by the University of Málaga**.**

Auto-adaptative Robot-aided Therapy based in 3D Virtual Tasks controlled by a Supervised and Dynamic Neuro-Fuzzy System

L. D. LLedo, A. Bertomeu, J. Díez, F. J. Badesa, R. Morales, J. M. Sabater, N. Garcia-Aracil

Miguel Hernández University of Elche, Elche, Spain

***Abstract* — This paper presents an application formed by a classification method based on the architecture of ART neural network (Adaptive Resonance Theory) and the Fuzzy Set Theory to classify physiological reactions in order to automatically and dynamically adapt a robot-assisted rehabilitation therapy to the patient needs, using a three-dimensional task in a virtual reality system. Firstly, the mathematical and structural model of the neuro-fuzzy classification method is described together with the signal and training data acquisition. Then, the virtual designed task with physics behavior and its development procedure are explained. Finally, the general architecture of the experimentation for the auto-adaptive therapy is presented using the classification method with the virtual reality exercise.**

***Keywords* — Rehabilitation robotics; Physiological state; Neural networks; Fuzzy logic system; Virtual reality; Collision detection.**

I. Introduction

THERE are a great number of literature about the growing importance of the use of robotic systems in the neurorehabilitation field [1] , [2]. Particularly in the assisted robotic devices for motor retraining in subjects who have suffered neurological injuries such as stroke or Parkinson. It is known that this kind of rehabilitation therapies produce a beneficial effects in those patients [3]. Many researchers intend to include them in a control loop [4] to increase the efficiency and effectiveness of such systems. Thus, robot-assisted systems are able to decide the difficulty level that can be made during the different rehabilitation therapies taking into account the emotional and physiological aspects of the subject.

Currently, the adaptation of the robotic systems behavior using psycho-physiological measures is analyzed by the scientific community. A large number of classification methods and emotional estimation are compared in [5]. However, there are few studies about the utilization of neuro-fuzzy methods in these subjects. The hypothesis that neural network help us to estimate the emotional state of the patients is supported due to the network theory can be applied into the neural computing of the emotions, as is described conceptually in [6], and the architecture of cognitive networks, affective networks and evaluation layers is proposed in [7].

Furthermore, the virtual reality is a technology that allows developing rehabilitation environments such as virtual therapies based on activities of daily living (ADL), intended for stroke patients [8]. In other research, it was compared the virtual rehabilitation with the classical rehabilitation with two different post stroke groups [9]. The result was that the group exposed to the virtual rehabilitation shown a better improvement in the motor deficits of the upper limb than the other group. For this reasons, the adaptive robot-assisted rehabilitation therapy can be beneficial for the patient.

This paper proposes a neuro-fuzzy architecture combined with 3D virtual reality in the development of an upper-limb rehabilitation application to study the potential usefulness of neural networks, fuzzy logic and three-dimensional environments based in physical principles. This method can be dynamically modify and can adapt the robot-assisted rehabilitation therapy according to the emotional state of the patient, following the psycho-physiological computing processing as defined in [5].

II. Classification Method

In this section, an analysis of the classification method, used to differentiate the emotional state of the user during the proposed robot-assisted rehabilitation therapy, the acquisition data process and the extraction features of physiological signals are explained. Further information about the learning algorithm and validation test is also presented.

A. Neuro-fuzzy System: S-dFasArt

S-dFasArt [10] is a classification method based on the architecture provided by the ART neural network [11], where fundamentals of the Fuzzy Sets theory are applied in the different processing stages of the classification algorithm. The neuro-fuzzy architecture takes the advantages of both techniques, the learning and adaptability capacity of the neural networks, and the robustness, the interpretability and the fault tolerance of the fuzzy systems.

DOI: 10.9781/ijimai.2015.328

To improve the convergence speed and the update mode of the fuzzy weights dynamical equations are used. Moreover, this classification method allows a fast, supervised and competitive learning, keeping the accumulated knowledge. All nodes or output categories are actuated by the input data due to its competitive property, but only the neuron with the highest response level is activated. Therefore, the winner category generates the classifier output on the current input pattern.

The architecture of the proposed method is shown in Fig. 1. This model is formed by three layers or levels and an orientation subsystem.

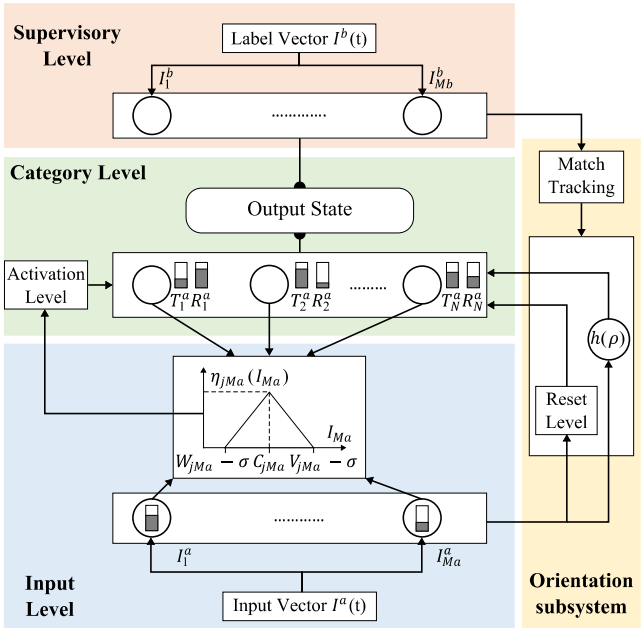


Fig. 1 S-dFasArt Architecture.

- The Input Level is used to receive the input data in a vector form, with the most important features of the physiological signals. Furthermore, in this level it is also applied an activation-membership function to determine the membership degree of each input attribute on the generated categories [12].
- The Supervisory Level presents the output pattern for the association with the input pattern.
- The Category Level is formed by a several nodes that represent the fuzzy units or categories and it contains the classification results with activation and reset levels. This level stores the association between the input sequences and its corresponding supervision vector.
- The Orientation Subsystem is responsible to create categories during the learning and the generalization of the similar categories.

B. Training data acquisition

To test the possibilities of the proposed classification method it is necessary the acquisition of the training data. The process to obtain the physiological signals is done as it is explained in [14] A robotic device designed for upper-limb assisted therapy called PUPArm, which is commercialized by Instead Technologies Inc, is used during the therapy. It has a signal acquisition system provided by g.tec medical engineering GmbH with different sensors and a virtual reality system. The physiological signal of the subject, such a pulse rate, respiration rate, skin conductance level (SCL), skin conductance response (SCR) and skin temperature [15], were recorded in real-time.To test the possibilities of the proposed classification method it is necessary the acquisition of the training data. The process to obtain the physiological signals is done as it is explained in [14]. A robotic device designed for upper-limb assisted therapy called PUPArm, which is commercialized by Instead Technologies Inc, is used during the therapy. It has a signal acquisition system provided by g.tec medical engineering GmbH with different sensors and a virtual reality system. The physiological signal of the subject, such a pulse rate, respiration rate, skin conductance level (SCL), skin conductance response (SCR) and skin temperature [15], were recorded in real-time.

The integrated virtual reality system encourages different psycho-physiological states of the patient. The activity is formed by a series of rectangular elements of different sizes that they move randomly across the screen with different speeds inside a defined area. Meanwhile, the user control a pointer with the robotic device in order to avoid the collision with the rectangular elements. Three different levels of difficulty (relax level, medium level and stress level) were defined using the number and speed of the rectangular elements shown in screen.

Once the physiological signals are acquired, a data processing, based on normalization of the features, was performed to get the final set of training data with its

respective supervision measures. The emotional states of the patient are collected in these supervision data.

C. Learning algorithm

In this section, the S-dFasArt neuro-fuzzy learning algorithm is explained briefly showing the required steps to train the neuro-fuzzy network nodes. Detailed explanation about this algorithm can be found in [13].

$$\frac{dT_j}{dt} = -A_T T_j + B_T \prod_{i=1}^{M} \eta_{ji}\left(I_i^a(t)\right) \tag{1}$$

$$\begin{aligned} l_{ji} &= max\left(V_{ji}, I_i^a\right) - min\left(W_{ji}, I_i^a\right) \\ \frac{dR_j}{dt} &= \left(-A_R R_j + B_R dreset\right)\left(R_{max} - R_j\right) \\ R_j(0) &= 0 \end{aligned} \tag{2}$$

The input training data and its corresponding supervision labels are received through the classifier at the Input Level and Supervisory Level respectively. Next, the activation level (1)

and the reset level (2) of all nodes that form the Category Level are calculated.

Then, the winner category is determined by comparing the activation levels values (3) and selecting the node with the highest value. If this activation value is null the classifier add a new uncommitted category.

$$T_J = max\{T_j; \quad j = 1...N\} \tag{3}$$

In the next step, it is necessary to check the reset level to determinate if the winner category accomplishes the necessary conditions of similarity with the input vector using a vigilance threshold. If it overtakes the threshold, the category whose level activation is the next highest, is searched. Once the winner category exceeds activation and reset conditions, its supervision label is compared with the supervision label received to confirm if it is a correct prediction. If it is not satisfied, the next category is sought.

To perform the update of the weights nodes of the neuro-fuzzy network the following dynamic equations (4) are used in case of an existing category.

$$\begin{aligned} \frac{dW}{dt} &= -A_W W + B_W min\left(I^a(t), W\right) \\ \frac{dC}{dt} &= A_C\left(I^a(t) - C\right) \\ \frac{dV}{dt} &= -A_V V + B_V max\left(I^a(t), V\right) \end{aligned} \tag{4}$$

In case of a new category, the weights nodes are initialized (5) using the input vector.

$$W = C = V = I^a(t) \tag{5}$$

D. Validation Test

At this point, a functional classification model has been implemented to distinguish correctly in real time the input patterns generated by the physiological signals of the patient and they are processed for features extraction. To get this functional model, an adjustment process of the SdFasArt neuro-fuzzy classifier have been applied following three phases [10]:

1. Initialization of the static value parameters linked to the dynamic equations.
2. Learning the weights whose values represent the diffuse categories generated by the presentation of the input pattern data to the classifier, using the learning algorithm explained in the previous subsection.
3. Setting the most influential parameters of the neuro-fuzzy network, σ and A_T, related to the diffuse character and the activation speed of the classifier categories, checking the values that provide better classification results.

Once the possible values of σ y A_{T} are analyzed, the functional classification model is completed. This model has been tested using the Leave-one-out validation technique to evaluate the classification results and ensure that they are independent of the partition between training and test data. Finally, the accuracy level has been calculated obtaining a performance results of 92.38% with 34 diffuse categories.

III. 3D Visual Tasks Architecture

The virtual task developed to perform the robot-assisted rehabilitation therapy has been designed following the scheme of programming routines. The general architecture is shown in Fig. 2. The graphic content required for the virtual scene is generated using Blender [16], a modeling software. This modeling tool can generate 3D meshes files with an appropriate format for an easy installation. Thus, the polygon mesh of the stage and the interacting elements are obtained. A file with the distribution of all scene elements is also generated.

These visual elements implemented are interpreted by the application core to build the virtual environment providing a characteristic behavior to each element. The core is responsible for controlling the application execution to organize all components of each of the blocks of the general scheme. The open source software called Ogre3D [17] has been used as a graphics engine. Ogre3D provides a flexible and object-oriented programming and, through its high level interface written in C++, it offers a series of intuitive methods that facilitates the preparation of 3D visualization applications with quite realistic, interactive and real-time environments on any kind of platform.

The physical engine is an important component that has been added. In this case, to simulate elements with some realism degree the NVIDIA PhysX \cite{physx} has been used. This engine tries to predict the physics effects within a scene subject to various conditions of speed, force, friction, mass and many more physics variables. Its main function is the collision detect algorithms able to calculate the interaction between all elements of the physics scene and the forces generation. Then, the rendering loop of the graphic engine updates the temporal evolution of the physics elements. It is also responsible for the visual representation of the behavior of each physics model by moving its corresponding graphical model. Each physics model consists in a simplified geometry of each specific graphical model to provide the main core the collision detect within the simulated physics space.

A sounds engine has been also incorporated to play sound effects depending on the task objectives. This engine helps in realism of the application and immerse the user in the virtual environment. The robotic system is controlled by the patient to manage the position of a virtual tool, whose movement affects in the collisions detect generated by the physics engine, depending of the simulated tool geometry. Meanwhile, the user has a visual feedback of everything happening in the application.

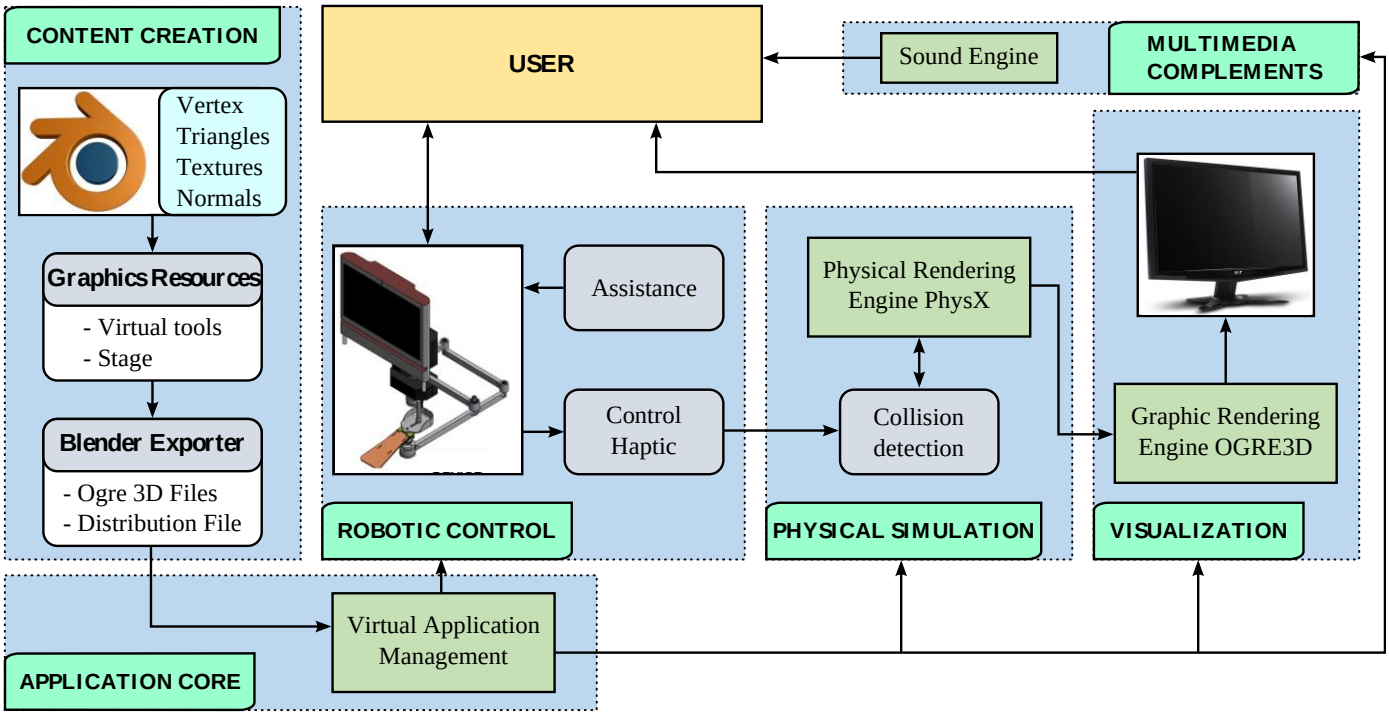


Fig. 2 General architecture of virtual reality applications.

A. Whack a Rabbit

In this paper, an example of virtual task has been designed using the procedure explained previously. The main objective of this task is to find items, in this case rabbits, which appear randomly on the scene. The user should use the robotic system to control the virtual tool and hit these targets. In Fig. 3 the implemented virtual task is shown.

To complete the experimentation different difficulty levels have been implemented in the virtual task for automatic level changes depending of the emotional state of the patient.

- Only one rabbit appears and stays on stage for 10 seconds.
- Only one rabbit appears and stays ion stage for 5 seconds.
- Two rabbits appear at the same time, and they stay on stage for 10 seconds.
- Two rabbits appear asynchronously, and everyone stays on stage for 5 seconds.
- Three rabbits appear at the same time, and they stay on stage for 10 seconds.
- Three rabbits appear asynchronously, and everyone stays on stage for 5 seconds.

The physic and visual elements used to develop the virtual task are explained in the following subsections.

Fig. 3 Virtual task screenshot

1) Static bodies

The static bodies are the elements of PhysX that are used in this application. These elements are placed in a particular localization in the scene and they are kept immobile during all simulation, while the remaining elements are influenced by the collision forces inside the stage during the exercise. In this virtual task the static elements have been designed to represent a garden with lawn, a wooden fence and 9 rabbit burrows. Also, a garden tools have been added to introduce the patient on the virtual reality.

2) Kinematic bodies

The kinematic bodies are objects that the user can freely

move within defined work range, and do not act under the influence of forces response, gravity and collisions. The movement of the kinematic bodies are produced getting the coordinates generated by the robotic system. In this task, a hammer has been designed as kinematic body, which generates an hit animation when the hammer is closed to the target.

3) Soft bodies

To provide more realism when the collision with the target occurs, the PhysX feature of soft bodies generating has been used. These deformable and volumetric elements have a elastic topology formed structurally by tetrahedral meshes that encapsulate all the 3D object surfaces. The stretching and bending constraints maintain subjected the vertices of this mesh type. The vertex positions are modified because of the internal forces of the soft body and the external forces generated in the physical engine.

Each tetrahedron of the soft body involves a number of vertices in the visual mesh. These vertices are attached to the behavior of the corresponding tetrahedron. Once the physics are updated, the new vertices positions are calculated and the graphic engines change the vertices positions of the visual mesh depending on the displacement suffered by the soft body. Thus, the 3D soft visualization body is updated. Also, PhysX offers different parameters that influence the internal physics of these bodies such as density, friction or damping, whose values can cause varying impact on the overall response of the element.

In this task, the bodies that move randomly in the scene are the soft bodies and they are shaped like a rabbit. Whenever a body is hit by the virtual tool, it color changes to indicate the success. Therefore, thanks to this functionality, the user is aware of the impact moment against the target and force direction.

IV. Autoadaptative Experimentation

The final application, based on adaptive robot-assisted rehabilitation therapy according with the specific psycho-physiological state of each patient, is defined in this section. A good performance has been observed in the classification method used in this experimentation, explained in the section II.

The general diagram of the autoadaptive experimentation is shown in Fig. 4, it is formed by four functional blocks. The sensors are connected to the user to record his/her physiological signals while controlling the interaction system. The explained rehabilitation robotic device and the virtual reality subsystem are included in this interaction system. The user has three feedback types (visual, audio and force). When the user starts using the robotic system, his physiological signals are processed to extract the most important features. These features are sent to the functional model of the neuro-fuzzy classifier. Three possible states are generated by the classification block processing, the input information to check the psycho-physiological state of the user and decide the modification to be made in the deployed task (up level, down level or maintain level). This proportionate classification state is sent to the interaction system to complete the experiment of automatic difficulty level changes in the virtual task explained in section II, where six levels have been defined. The aim of the experimentation is to maintain an intermediate difficulty level along the task period.

The extraction and processing of the signals features have been performed in a Simulink scheme, while the classification block has been designed with Matlab. A UDP protocol is used to communicate the classifier with the virtual task software. This communication is applied every 30 seconds to send three possible action commands and changing the difficulty level of the virtual task:

1. If the user has an over-stressed level, one difficulty level is reduced in the virtual task.
2. If the user has stable state, the virtual task does not change the difficulty level.
3. If the user is relaxed, the actual difficulty level is increased.

Currently, several subjects are doing these experiments, who were informed of the work purpose. The subjects must perform an adaptation period of few minutes. Then, the subjects signals in relaxed state for 5 min are recorded to obtain baseline measurements. Finally, the subjects perform the task for 10 minutes, starting in the first difficulty level of the virtual task.

V. Conclussion

In this paper, a classification model based on a neural network architecture and the fuzzy logic is presented to classify three different states of the user's reactions physiological in an assisted system of robotic rehabilitation. This method has obtained a quite good performance results (92.38% in LOOCV) that allow an efficient classification in real-time.

On the other side, a graphical application based on 3D virtual reality has been implemented. The aim of this issue is to increase the sense immersion of the user inside of the virtual environment, checking the patients behavior in front of scenes in three dimensions. Ogre3D has been used for visualization. The physics engine adds realism to the scene providing performance energies on elements that have different physical behaviors. These software tools offer a great versatility and flexibility to implement all type of virtual exercises with visually and physically realistic environments.

Acknowledgment

This work has been supported by the Spanish Government through the project ``Interpretation of the human intention and performance through biomedical signals and analysis of kinematic and dynamic motion" (DPI2011-29660-C04-04), by CIBER BBN and by MAT2012-39290-C02-01.

REFERENCES

[1] E. L. Miller, L. Murray, L. Richards, R. D. Zorowitz, T. Bakas, P. Clark , S. A. Billinger, "On behalf of the American Heart Association Council on Cardiovascular Nursing, and the Stroke Council. Comprehensive overview of nursing and interdisciplinary rehabilitation care of the stroke patient: A scientific statement from the American Heart Association", Stroke 41(10), pp. 2402-2448, 2010.

[2] "The Management of Stroke Rehabilitation Working Group on behalf of the Department of Veterans Affairs, Department of Defense, and The American Heart Association/American Stroke Association, VA/DoD Clinical Practice Guideline for the Management of Stroke Rehabilitation". Available: www.healthquality.va.gov. Accessed 25 Feb 2014.

[3] N. Norouzi-Gheidari, P. S. Archambault, J. Fung, "Effects of robot-assisted therapy on stroke rehabilitation in upper limbs: systematic review and meta-analysis of the literature",Journal of Rehabilitation Research and Development 49(4), pp. 479496, 2012.

[4] A. Duschau-Wicke, A. Caprez, R. Riener, "Patient-cooperative control increases active participation of individuals with SCI during robot-aided gait training", Journal of NeuroEngineering and Rehabilitation 7(43), 2010.

[5] D. Novak , M. Mihelj, M. Munih, "A survey of methods for data fusion and system adaptation using autonomic nervous system responses in physiological computing", Interacting with Computers 24(3), pp. 154-172, 2012.

[6] L. Pessoa, "On the relationship between emotion and cognition",Nature Reviews Neuroscience 9(2), pp. 148-158, 2008.

[7] F. Shu, A. H. Tan, "A biologically-inspired affective model based on cognitive situational appraisal", Neural Networks (IJCNN), The 2012 In- ternational Joint Conference, doi: 10.1109/IJCNN.2012.6252463, 2012.

[8] K. Laver, S. George, S. Thomas, J. Deutsch, M. Crotty, "Virtual reality for stroke rehabilitation", Cochrane Database Syst Rev 9:CD008349, 2011.

[9] A. Turolla, M. Dam, L. Ventura, P. Tonin, M. Agostini, C. Zucconi, P. Kiper, A. Cagnin, L. Piron, "Virtual reality for the rehabilitation of the upper limb motor function after stroke: a prospective controlled trial", Journal of NeuroEngineering and Rehabilitation, pp.10:85, 2013.

[10] J. M. Cano-Izquierdo, J. Ibarrola, M. Almonacid M, "Improving Motor Imagery Classification with a new BCI Design using Neuro-Fuzzy SdFasArt", IEEE Transactions on Neural Systems and Rehabilitation Engineering 10(1), pp. 2-7, 2012

[11] G. Carpenter, S. Grossberg, N. Markuzon, N. Reynolds, D. Rosen, "Fuzzy ARTMAP: A neural network architecture for incremental supervised learning of analog multidimensional maps", IEEE Transactions on Neural Networks 3(4), pp. 698-713, 1992.

[12] J. M. Cano-Izquierdo, Y. A. Dimitriadis, E. Gomez-Sanchez, J. Lopez-Coronado, "Learning from noisy information in FasArt and FasBack neuro-fuzzy systems", Neural Networks 14(4-5), pp. 407-425, 2001.

[13] J. M. Cano-Izquierdo, M. Almonacid, M. Pinzolas, J. Ibarrola, "dFasArt: Dynamic neural processing in FasArt model", Neural Net- works 22(4), pp. 479-487, 2009.

[14] F. J. Badesa, R. Morales, N. Garcia-Aracil, J. M. Sabater, A. Casals, L. Zollo, "Auto-adaptive robot-aided therapy using machine learning techniques", Computer Methods and Programs in Biomedicine 116(2), pp. 123-130, 2014

[15] F. J. Badesa, R. Morales, N. Garcia Aracil, J. M. Sabater, C. Perez-Vidal, E. Fernandez, "Multimodal interfaces to improve therapeutic outcomes in robot-assisted rehabilitation", IEEETransactions on Systems, Man, and Cybernetics, Part C 42(6), pp. 11521158, 2012.

[16] Gonzalez-Crespo, R., S. Rios-Aguilar, R. Ferro-Escobar, and N. Torres, "Dynamic, ecological, accessible and 3D Virtual Worlds-based Libraries using OpenSim and Sloodle along with mobile location and NFC for checking in", IJIMAI, vol. 1, issue Regular Issue, no. 7, pp. 62-69, 12/2012

Luis Daniel Lledó received telecommunications engineering from the Miguel Hernández University of Elche, Elche, Spain, in 2010, where he is currently working towards the Ph.D. degree. His research interests include the development of virtual reality applications for rehabilitation tasks, surgical simulation and neuro-fuzzy systems.

Arturo Bertomeu received the telecomunication engineering degree from the Miguel Hernández University of Elche, Elche, Spain, in 2014, where he is currently working toward the Ph.D. degree. His research interest include rehabilitation robotics, control systems and kinematic reconstruction of the human body limbs.

Jorge Díez received the industrial engineering degree from the Miguel Hernandez University, Elche, Spain, in 2014, where he is currently working toward the Doctorate degree in the field of rehabilitation robotics at Biomedical Neuroengineering Group with Prof. N. Garcia-Aracil. His research interests include industrial and mechanical design of rehabilitation and assistance robots, control systems, and CAD programming.

Francisco Javier Badesa received the M.Sc. degree in telecommunications engineering in 2008, and the Ph.D. degree in telecommunications and industrial technologies in 2014, both from the Miguel Hernández University of Elche (UMH), Elche, Spain. His research interests include rehabilitation robotics, control systems, the development of virtual reality environments for therapeutical purposes, and biomedical signal processing.

Ricardo Morales received the degree in industrial engineering, in 2004, and the Ph.D. degree in telecommunications and industrial technologies in 2014, both from the Miguel Hernández University of Elche (UMH), Elche, Spain. His research interests include robotics for stroke rehabilitation, control systems, and safe Physical Human-Robot Interaction.

José María Sabater received the M.Sc. degree in nuclear engineering from the Polytechnic University of Valencia (UPV), Valencia, Spain, in 1998, and the Ph.D. degree in industrial engineering from Miguel Hernández University of Elche (UMH), Elche, Spain, in 2003.
He is currently an Associate Professor of Control and Systems Engineering with UMH. In 1998, he was an Industrial Engineer with the UPV.

Nicolas Garcia-Aracil is Associate Professor of Control and Systems Engineering at Miguel Hernandez University (Spain). He holds a M.Sc. in Control Engineering by the University of Murcia (1996, Spain), Master in Design, Robotics and Industrial Automation from University of Murcia (Spain) 1996-1997 and a PhD in Control Engineering by the Miguel Hernandez University of Elche (Spain). His current research interests are medical and surgical robotics, rehabilitation robotics, medical image, computer vision, human-robot Interaction and design and control of new robotic devices. Dr. Nicolas Garcia was a 2004 recipient of the Best Thesis in Roboticsˇ, National Research Prize, from the Spanish Federation of Automatic control. He is author or co-author of a broad range of research publications. He served as Program Chair of the 2012 IEEE RAS\\EMBS International Conference on Biomedical Robotics and Biomechatronics (BIOROB, Roma, Italy) and the General Chair of EURON Winter School on Rehabilitation Robotics (Elche, Spain).

GPGPU Implementation of a Genetic Algorithm for Stereo Refinement

Álvaro Arranz, Manuel Alvar

Zed Worldwide

***Abstract* — During the last decade, the general-purpose computing on graphics processing units Graphics (GPGPU) has turned out to be a useful tool for speeding up many scientific calculations. Computer vision is known to be one of the fields with more penetration of these new techniques. This paper explores the advantages of using GPGPU implementation to speedup a genetic algorithm used for stereo refinement. The main contribution of this paper is analyzing which genetic operators take advantage of a parallel approach and the description of an efficient state- of-the-art implementation for each one. As a result, speed-ups close to x80 can be achieved, demonstrating to be the only way of achieving close to real-time performance.**

***Keywords* — Parallel processing, GPGPU, genetic algorithm, stereo.**

I. Introduction

RECENTLY, custom GPU programming has become one of the most popular tools for increasing the efficiency of parallel algorithms thanks to the computational capacity of the Graphics Processing Unit (GPU) compared to serial CPU programs.

Traditionally, GPUs appeared in the computer market as hardware products specialized on rendering tasks and, more specifically, for improving the gaming experience. Given that most of the rendering pipeline's steps of were parallel, these products rapidly evolved to machines capable of efficiently running highly parallel algorithms. In the last decade, the flexibilization of the GPU hardware and tools has enabled the use of these parallel-processing units for general scientific purposes.

Stereo analysis is a Computer Vision research area that has been widely studied in the literature. However, it remains an unsolved problem and many algorithms are still proposed every year. The aim of the stereo analysis is to obtain depth information from a couple of stereo images, simulating how the human's can perceive the depth using just two eyes. Solving this problem is very computationally demanding, especially when dealing with high-resolution images. GPGPU techniques have been recently used for speeding up these tasks and great results have been reported in the literature.

GPGPU primarily aims to improve the program's performance. It has been demonstrated that using these techniques could result in a speed-up of up to x100, depending on the algorithms' nature. This paper proposes to study the speed-up achieved by GPGPU programming applied to an evolutionary algorithm. A genetic algorithm for stereo refinement is implemented in both CPU and GPU and its performance analyzed and compared.

Improving the accuracy and performance of stereo algorithms is crucial for many real applications. Robotics has been traditionally a research area that has used these techniques, but new fields are arising. The digitalization of the automotive sector is leading to the incorporation of new sensors such as high definition cameras to high-end cars. Fast stereo algorithms are needed to provide accurate information about the car's environment. Other applications of stereo algorithms are biomedicine, virtual reality, automation or the entertainment industry. However, note that any optimization problem solved with evolutionary algorithms might benefit from the work herein proposed.

The paper is structured as follows: Section II is a brief overview of the GPGPU implementations found in the literature, Section III explains the stereo refinement genetic algorithm implemented, Section IV describes the details about the GPGPU implementation, in Section V some results are presented and finally in Section VI some conclusions are drawn.

II. GPGPU Overview

GPGPU has been widely used in the literature by the computer vision community. Its main role has been to enable real-time performance on many demanding algorithms.

First works on stereo GPU processing were proposed in [11]. SSD dissimilarity techniques, a multi-resolution approach and a very primitive GeForce4 were used to obtain performance equivalent to the fastest CPU commercial implementations available. Later, [3] proposed a multi-view plane- sweep-based stereo algorithm for handling correctly slanted surfaces applied to urban environments. Assuming a highly structured scene with buildings, they used a planar prior for estimating disparity maps. The algorithm was successfully implemented in an Nvidia GPU obtaining real-time frame rates.

DOI: 10.9781/ijimai.2015.329

In [6] a high-performance stereo-matching algorithm both fast and accurate is proposed. Using a parallel designed AD-census and scanline optimization implemented in CUDA in an NVIDIA GeForce GTX 480 they achieve near to real- time frame-rates. They report an impressive 140x speed up compared to the CPU implementation. Another GPU stereo matching algorithm using adaptive windows can be found in [13].

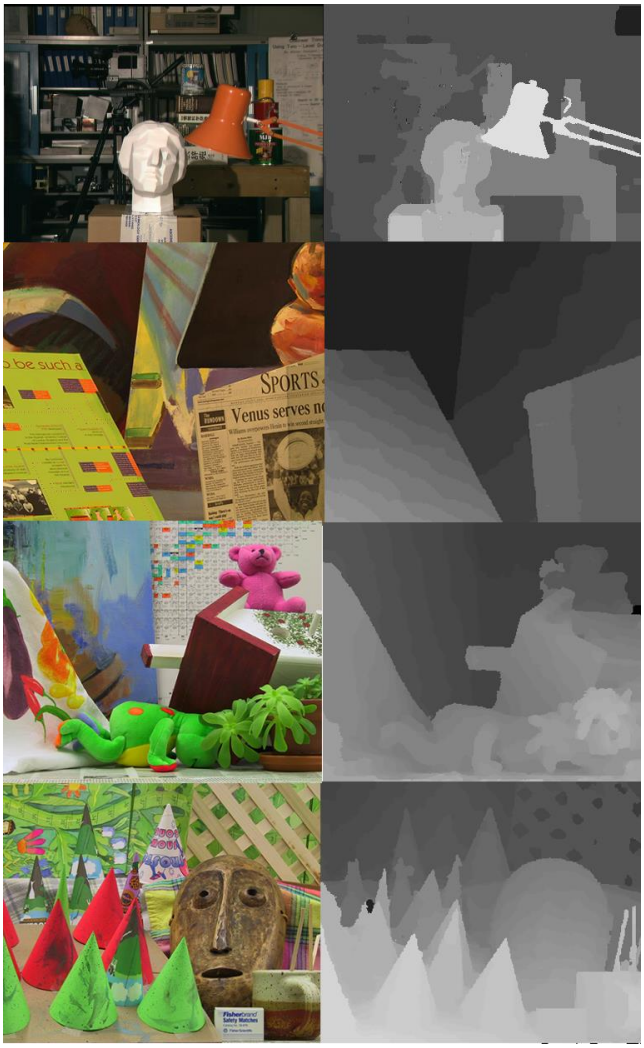


Fig. 1 Disparity map examples

In [4] and [7] a real-time camera tracking and mapping using RGB-D cameras is proposed, obtaining quite impressive results. Their implementation relays heavily on the use of GPGPU, both for tracking and TSDF mapping. Depending on the voxel's resolution, they achieve execution times from 10 to 25ms.

GPGPU has also been applied in other fields, such as in feature detection and tracking, as proposed in [9]. Their KLT GPU implementation achieves real-time 30Hz on 1024768 resolution images, which is a 20x speed-up compared to their CPU implementation. A 10x improve is also reported for the SIFT [5] detector implemented in GPU. A CUDA implementation of the famous graph cuts algorithm [2, 14, 15] is presented in [10], obtaining a 12x performance enhancement.

A similar system to the one herein proposed is presented in [8]. In this work, a genetic algorithm for stereo matching is also implemented in GPU. However, the genetic algorithms have quite different approaches, and their parallel implementation does not seam to provide any performance boost compared to the CPU one. This paper shows that, with the proper GPU implementation, a 50x speed-up can be achieved.

III. Genetic Algorithm for Stereo Refinement

The genetic algorithm for stereo refinement implemented in this paper is based on the work proposed in [1]. The implementation minimizes a fitness function that is related to a Markov Random Field (MRF) and is equivalent to minimizing a global energy function. Due to the flexibility of genetic algorithms, this function is able to include occlusion handling.

This algorithm uses a guided search approach with new crossover and mutation operators adapted to the stereo refinement problem. Each operator will be explained briefly in this section. An example of the results that can be achieved using these techniques is shown in Figure 1.

A. Genome representation

Each individual includes the whole disparity map estimate and the occlusion map for both left and right images.

$$\bar{g}\begin{cases}\bar{f}_L=\{f_{L_1},f_{L_2},\ldots,f_{L_N}\}\\ \bar{f}_R=\{f_{R_1},f_{R_2},\ldots,f_{R_N}\}\\ \bar{O}_L\\ \bar{O}_R\end{cases},\quad f_i\in\Lambda,\Lambda=\{1,2,\ldots,L\}\qquad(1)$$

where g is the genome, gL and gR are the representation of the left and right disparity images respectively, XiL and XiR are the disparities estimated for pixel i on the left and right disparity images, N the total number of pixels in each image and Li the set of different disparity labels.

Occlusion maps are defined as:

$$O(p)=\begin{cases}0 & \textit{if not occluded}\\ 1 & \textit{if occluded}\end{cases}\qquad(2)$$

where O(p) is the occlusion map and p is the pixel.

B. Initialization

For the initialization process two different window-based algorithms with different window sizes, the adaptive support-weight approach [12] with random parameters and the census based with window-cost aggregation have been used. This variation aims to provide a wide range of initial solutions.

$$E(\bar{g}) = E_{data}(\bar{g_L}) + E_{smooth}(\bar{g_L}) \quad (3)$$

$$E_{data}(\bar{g_L}) = \begin{cases} \lambda_d & \text{if } i \text{ is occluded} \\ \sum_{i \in \bar{g_L}} |I_L(x_i, y_i) - I_R(x_i - X_i, y_i))| & \text{otherwise} \end{cases} \quad (4)$$

$$E_{smooth}(\bar{g_L}) = \sum_{\{p,q\} \in N} min\left(\frac{\beta_s}{\varphi_s}|X_p - X_q|, \lambda_{st}\right) \quad (5)$$

$$\beta_s = max(\lambda_s, \gamma_s - |I_L(p) - I_L(q)|) \quad (6)$$

C. Fitness function

An energy function that considers discontinuities and occlusions is used for the fitness function:

where g is a certain individual, gL is the left disparity image, Il and Ir stand for the left and right stereo pair, xi and yi are the image coordinates of pixel i, V{p,q} is a smoothing function and λs, γs and φs are constant parameters for every pixel.

Before any fitness function evaluation, a occlusion management process is triggered for classifying pixels correctly before any energy evaluation.

D. Occlusion management

The process of handling the occluded areas is a two-step operation: occlusion detection followed by an occlusion management.

The following operations are defined for calculating the left occlusion map:

$$O_L(p) = \begin{cases} 0 & \exists i / \begin{pmatrix} x(i) + \bar{g_R}(i) \\ y(i) \end{pmatrix} = \begin{pmatrix} x(p) \\ y(p) \end{pmatrix} \quad p, i \in P \\ 1 & \text{otherwise} \end{cases} \quad (7)$$

being OL the left occlusion map, x(p) and y(p) the x and y coordinates of point p respectively and P the set of disparity image points. A similar expression can be deduced for the right occlusion map. This occlusion map identifies which areas of the image are classified as occluded regions.

For the occlusion management, an iterative process based on neighboring disparities of the occluded pixels is applied. For the left image, each occluded pixel is assigned the disparity value of the most photo-consistent non-occluded neighbor from left to right and afterwards it is marked as non-occluded. If no non-occluded neighbors exist, it maintains its occluded status for the next iteration. Special status have the occluded pixels whose x(p) coordinate is less than the number of disparities analyzed. In this case the iteration is made from right to left and bottom-up. The iteration is finished when no occluded pixels are left on the left occluded map.

For the right image it is similarly done but vice versa (right to left for common pixels and left to right for pixels whose x(p) is at a distance of the number of disparities analyzed from the right image border).

E. Crossover

The crossover is based on comparing parent's blocks of different sizes and assign the best ones to the same son. This operator can be summarized in the following steps:

1) Parents are divided into blocks (random sizes)
2) The fitness function of each block is evaluated
3) Best block is selected to persist in the same child

F. Mutation

Three different mutation operations may occur to each individual. Firstly, one possible mutation operation is to initialize again a group of pixels following the steps explained in Subsection III-B with a probability PMa. Secondly, a bilateral filter operation with a random window size with a probability PMb. Finally, a morphological operation such as erode or dilate may occur with a probability PMc.

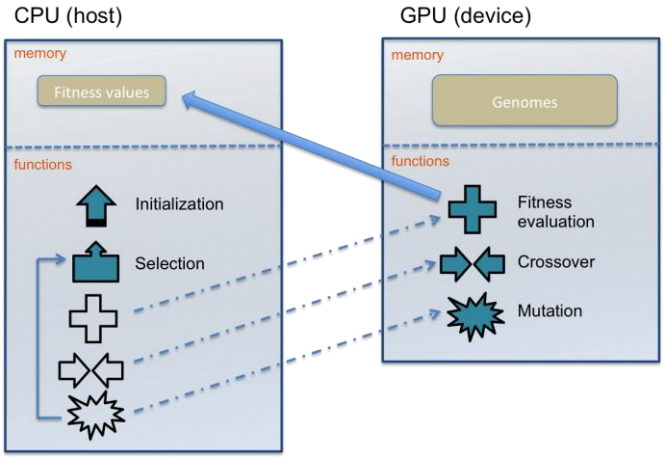


Fig. 2: Assignment of genetic operators to GPU and CPU

IV. Algorithm's GPGPU Implementation

After analyzing the performance of the serial version of the genetic algorithm, it is easy to conclude that the most computationally demanding functions are the genetic operators and not the genetic algorithm itself. This result is straightforward because each genome includes a lot of data and information inside (whole four images: two disparity maps and two occlusion maps). For example, each genome evaluation implies evaluating the energy function for each pixel and neighborhood individually. Besides, each genome operator is naturally parallel, which suggests that implementing these operators in CUDA will have a dramatic impact on the genetic algorithm performance.

In Figure 2 is shown where is computed each genetic operation. The left side of Figure 2 represents data information is stored and which functions are implemented and executed in the CPU. The fitness values are stored in the CPU because they are needed for the selection operator in order to decide which individuals of the actual population will survive to the next one. The right side of the diagram represents which information is stored and which functions are evaluated in the

GPU. All the genomes are stored in the GPU in order to enable fast access to the data from the functions evaluated in the device. The only memory transaction between the CPU and GPU needed is the copy of the fitness value of each individual from device to host and is represented by the big blue arrow from the fitness evaluation function icon to the fitness value memory in the CPU. Remember that this device-host and vice-versa transactions are very costly and must be minimized for achieving the best performance.

The genetic algorithm has been implemented in the CPU using the GAlib library. For the image processing and allocation it has been used the OpenCV library, specifically the GPU module, which facilitates the memory allocation and transaction and has quite a lot processing algorithms built-in in the GPU already. Finally, evaluation, crossover and mutation operators have been implemented in CUDA in several kernels. The next sections describe in detail the strategy used for implementing efficiently each operator in CUDA language.

1) CUDA evaluation kernel

Although the title may suggest that the evaluation of a genome is carried out just by one kernel, the reality is that it is a process composed by three steps. The first two are solved using a single kernel each while the third has to be solved by two kernels. The first two steps could be executed in parallel by two different CUDA streams but the last have to be executed after the firsts have finished. This parallel capability has not been implemented and all four kernels have been

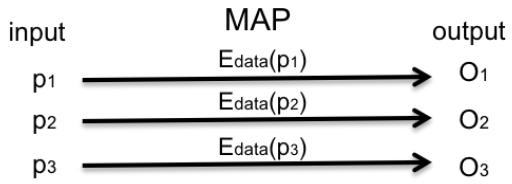


Fig. 3: Parallel MAP operation

programmed to run in the same stream.

The first step in the evaluation process is the data term evaluation of the energy function. The result is one value for each pixel and its calculation is independent from the values of the neighboring pixels. Thus, the relation is one to one and its parallel implementation is very efficient and straightforward. This type of operation is also called MAP, and it has been implemented using one thread per pixel in the disparity image. A simple diagram of MAP is shown in Figure 3.

The data term only depends on the values of the left and right stereo image and on the disparity image evaluated. Left and right stereo images have been allocated in the device as 2D textures, which are very efficient for interpolation. Note that in this case, using shared memory does not make much sense because the number of memory accesses needed per thread would not bee minimized. The result is saved in a floating-point structure of the same size as the original image, and here will be referred as memData.

The second step is the evaluation of the smoothing term. If one thread per pixel is used, it requires to access to its own disparity value and the neighbouring disparities. This operation can be considered a type of Stencil operation, in which many reads are needed as input while only one write is performed. An illustrative example is shown in Figure 4.

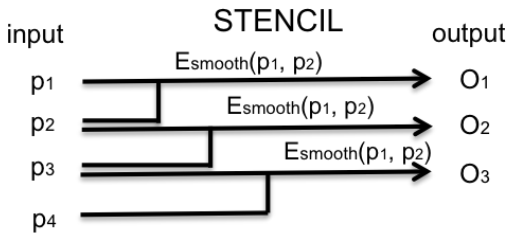


Fig. 4: Parallel STENCIL operation

In order to maximize the performance, shared memory is used for first loading all the disparities in a block and then using that shared memory in all the threads of the same block. Remember that the access to shared memory is much faster than the access to global memory. Each thread is related to each pixel in the disparity image and it is in charge of evaluating the smoothing function that relates itself with the right and bottom neighbors. As happened in the data kernel, the result is saved into a floating point vector with the same size of the disparity image and here will be called memSmooth.

The third and last step is composed of two kernels, one executed after the other. It is in charge of performing the summation over all pixels of the memData and memSmooth structures calculated in the two first steps. This type of summation is an operation also known as Reduce. Although at first glance this operation might seem difficult to parallelize, actually it is fairly simple. Figure 5 shows the two-step reduction implemented. Besides, this step sums memData and memSmooth individually for each pixel and saving it in memTotal in order to facilitate the crossover task explained in subsection IV-2.

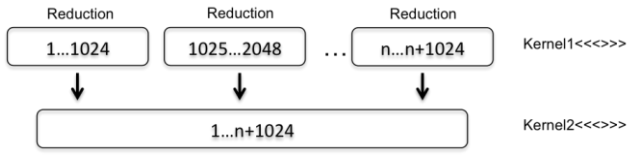


Fig. 5: Reduction operation in two kernels

For enhancing the performance of the Reduce operation, the data has been divided in groups of 1024 addends, each being processed by a CUDA block. All the data in each group is loaded in shared memory to improve its read and write speed. The first kernel performs the summation over each group, obtaining one result per group. Finally, the last kernel performs the last summation over all the results of the previous kernel, and obtains the final value for the fitness function.

Finally, an asynchronous memory copy is performed from device to host to copy the final fitness value calculated for that genome. This operation is recommended to be asynchronous because the memory copy can be performed at the same time

as other kernels are executed in other streams, instead of waiting until the memory copy has finished.

Note that in spite of running all the evaluation kernels in the same stream, different individuals are able to run their evaluation in different streams, which enables copying memory from device to host at the same time as other kernels and a higher level of parallel exploitation.

The evaluation process is performed over the left and right disparity map, but for the right one, it is not necessary to perform the final Reduction and memory copy. This optimization can be achieved because de fitness function of a genome is just the fitness function of its left disparity image.

2) *CUDA crossover kernel*

As explained in Subsection III-E, the crossover operation consists of three steps:

• Divide the two disparity maps into blocks. In this case, the number of pixel of each block will not be greater than 1024.

• Sum up the memTotal for each pixel inside the blocks, compare them by pairs (one for each parent) and keep the block with the best fitness function.

• Copy the best block to the children.

The limit of 1024 pixels per block is related to the maximum number of threads per CUDA block available by the GPU. All the pixels in a block must be part of the same CUDA block be- cause the summation can be performed using shared memory, which is much more efficient than global memory. Therefore, the CPU realizes the first step, and the following two are done by the GPU, the first one as a Reduction operation very similar to the one in subsection IV-1 and the second one as a very simple copy operation as a MAP.

3) *CUDA occlusion handling kernel*

Occlusion handling encompasses two different tasks: occlusion estimation and occlusion management. The occlusion estimation is calculated through an image warping, where each pixel of the other disparity image is displaced a number of pixels equal to its disparity level. Pixels left without any assignation are considered to be occluded pixels. Thus, each pixel operation is independent from the rest, but several threads can output their result to the same piece of memory. This operation is also known as Scatter and can be solved using, for example, atomic operations. In our case it is not necessary because the function aims only to output a boolean value, more precisely a zero to indicate that the pixel is not occluded.

The second task is the occlusion management, where the main objective is to re-estimate the disparity value for the pixels that where labelled as occluded. For this parallel implementation the horizontal fast occlusion filling algorithm explained in Subsection III-D was used. Given that a horizontal search for the closest non-occluded pixel has to be performed, the occlusion information was loaded in shared memory, being each block responsible for each independent scan-line. Each thread is in charge of estimating the new depth for each occluded pixel. Figure 8 shows the per-thread operations and the memory accesses incurred.

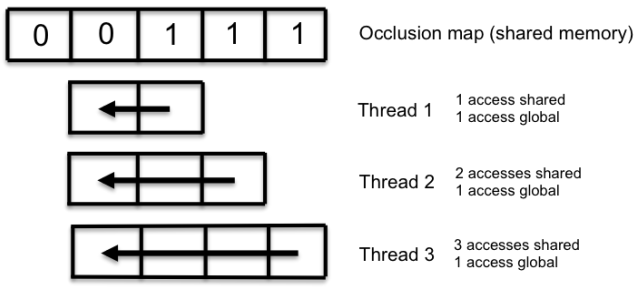


Fig. 6: Horizontal fast occlusion filling implementation example

4) *CUDA mutation kernel*

The mutation kernel comprises three different operations: bilateral filtering, erosion and dilation. These morphological operations are already efficiently implemented in the OpenCV library using CUDA. A problem that may rise using a third party library is the performance penalty incurred while parsing from the data-types used in your application to the data-types used in the library and vice- versa. However, in the implementation herein proposed, the data types are compatible with those from OpenCV, so this transformation is trivial. Thus, this library has been used for this purpose.

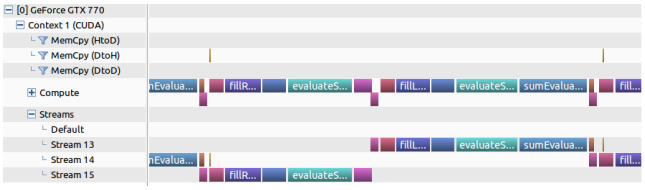


Fig. 7: Nvidia Visual Profiler tool

V. RESULTS

In this section the parallel capabilities of the genetic algorithm are discussed. Both the serial implementation and the parallel one using CUDA are compared. Given the stochastic nature of the algorithm and the various types of mutations that are likely to happen, the algorithm was run for different images during five hundred generations and an average per individual and generations was calculated. The Middlebury dataset will be used for comparison, as it is a standard and well-known test-bed.

For the tests, an Intel i7-2600 at 3.4 GHz CPU and an Nvidia GeForce GTX 770 were used. As operating system, Ubuntu Linux 14.04LTS was used given the CUDA performance improvement compared to Windows. The measuring tool used was the Nvidia Visual Profiler, obtaining valuable data such as timing, occupancy, optimizations, et. A capture of the profiler is shown in Figure 7.

The parameters used in the experiments carried out along this section are shown in Table I and Table II.

A comparison between the performances of the GPGPU versus the CPU implementation for four Middlebury's common test images is shown in Table III.

TABLE I
PARAMETERS FOR THE NEW ENERGY FUNCTION

λ_d	λ_s	γ_s	φ_s	λ_{st}
10.0	50.0	2.0	10.0	$ndisp/2$

TABLE II
PARAMETERS FOR THE GENETIC ALGORITHM

Population	Generations	P_{cross}	P_{Ma}	P_{Mb}	P_{Mc}
50.0	500	0.9	0.01	0.033	0.066

TABLE III
MEAN TIME SPENT FOR EACH INDIVIDUAL AND GENERATION IN CPU AND CUDA IMPLEMENTATION

	CPU (ms)	CUDA(ms)	Speed-up
Tsukuba	20.84	0.359	58.05
Venus	31.6	0.52	60.77
Teddy	45.63	0.579	78.8
Cones	46.77	0.574	81.48

The first column of Table III shows the mean total time spent for the CPU implementation for one genome. Note that not all genetic operations always occur in each genome and, therefore, these results are obtained dividing the total time spent by the algorithm by the number of genomes and generations. The increment in the CPU execution time from Tsukuba to Cones is explained due to the increment in the size of the test images.

The second row shows the same measure, but using now the GPU implementation. It is shown that parallelization of the genetic algorithm provides a great performance improvement compared with the serial one. The speed-up comparison between the two algorithms is shown in the third column. Note the increment in the performance improvement when the images get bigger, suggesting that with more pixels the GPU performs more efficiently. However, both CPU and GPU implementation still depend highly on the number of pixels in the image analyzed.

TABLE IV
MEAN TIME SPENT BY EACH GENETIC OPERATION FOR EACH INDIVIDUAL AND GENERATION IN TSKUBA AND CUDA IMPLEMENTATION

	time (ms)
Evaluation	0.155
Crossover	0.061
Occlusion handling	0.062
mutation	0.033
CPU	0.0467
Total	0.359

In order to study the impact of each genetic operation, Table IV shows in detail how the time is divided for each genome. It shows that evaluating the genome is the most demanding operation. Given that it is an operation that has to be run always in every gnome and that it is quite complex (energy function composed by several complex terms), this result is comprehensible. In comparison, the other operation that is run always and has a lot less impact in the total time is the occlusion handling. The percentage of the impact is shown in Figure 8.

As a result, it can be said that adding the occlusion handling to the algorithm implies a 17% impact on the performance. This result does not account for the impact of the occlusion variable in the evaluation operator, which here will be considered negligible.

TABLE V
MEAN TIME FOR EACH OPERATOR IN TSKUBA AND CUDA IMPLEMENTATION

	time (ms)	times per genome
Evaluation	0.155	1
Crossover	0.0663	0.9018
Occlusion handling	0.062	1
mutation	0.277	0.121

TABLE VI
MUTATION OPERATION EXECUTION TIMES

	time (ms)	times called (%)
Bilateral filter	0.273	36.78
dilate	0.283	32.19
erode	0.275	31.03

Maybe the result that was unexpected was the efficiency of the crossover function. However, although being a demanding operation, a lot of information from the evaluation process could be reused, leading to an efficient implementation. Bear in mind that the crossover it is run with a probability of Pcross, so this fact also has an impact on this measure. The same occurs with the mutation operation, that it is has a low impact due to it is rarely run.

Finally, a CPU entry in this table might seem strange at first. This time is attributed to the tasks of launching the CUDA kernels and managing the genetic algorithm itself, not the operators. As shown in Figure 2, this includes the selection operation, sorting, etc.

The measures presented in Table IV were calculated aggregating the occurrences of all the operations, but they do not occur in the same proportion. Therefore, those metrics do not represent the true performance penalty of each operation. In Table V the performance of each individual operator is shown.

These measures are the mean time spent value for each operation individually. It can be seen that, although the mutation operation has little impact on the total time spent on the algorithm, individually, it is by far the most demanding one. This is explained by the fact that a low mutation probability was set. Incrementing the mutation probability would have a great impact in the algorithm's performance. The second row of Table V shows statistically how many times each operation is called for each genome.

Finally, a more in-depth analysis of the mutation operation

is shown in Table VI.

The three different operations were configured to be triggered with the same probability, and this is represented in the second row of the table. It is shown that the three algorithms perform very similarly.

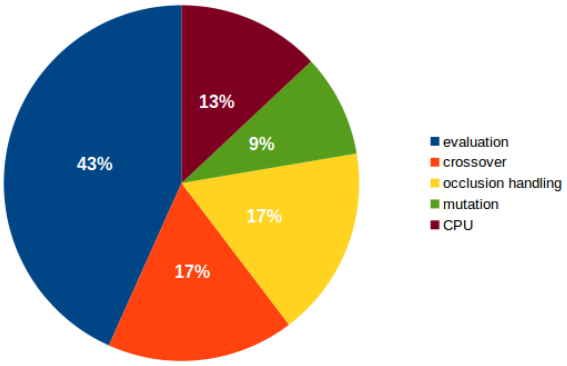


Fig. 8: Portion of time spent by each operation for Tsukuba

As a conclusion, it can be stated that approximately a 80x speedup can be achieved using a parallel implementation of the algorithm used on sufficiently big images.

VI. Conclusions

In this paper, a parallel GPGPU implementation of a genetic algorithm has been proposed. These evolutionary algorithms are very flexible and fit nicely in a parallel architecture given that the operators act independently on each individual genome. This quality suggests that parallelizing the main genetic operators would have a great impact in the algorithm's performance.

The most time-demanding genetic operators considered to be run in GPGPU were the fitness evaluation, the crossover and the mutation. However, the selection and the maintenance of the genetic algorithm itself was decided to be kept in the CPU. The main reason was that these tasks are negligible compared to the other operators; this assumption was supported by the results presented. Each operator was analyzed and a specific parallel implementation was proposed for each one.

A genetic stereo refinement algorithm with occlusion handling was selected for analysis. Using the standard Middlebury's stereo test-set, a comparison between a CPU and a GPGPU implementation was shown. As a conclusion, a great performance improvement can be achieved using GPGPU computation: a x80 speed-up has been achieved for some images. An analysis of the time spent by each operation and the impact of modifying the genetic parameters has been discussed. As a result, the most demanding operation was the fitness evaluation. This is reasonable due to the complexity of the energy function used for testing. However, considering individual function performance, the mutation operations are the most expensive, so an increment in the mutation probability would have a noticeable impact on the performance.

Evolutionary algorithms are generally not designed for real-time applications. Although a great performance improvement has been obtained, real-time performance is still not achievable for these applications. However, the GPGPU implementation improved the algorithm's performance from minutes to seconds order of magnitude.

In order to continue with this line of research, in future works, it would be interesting to try different genetic algorithm's formulations such as migrating or overlapping populations. These approaches might help avoiding local minima during the optimization. In [1] was demonstrated that this algorithm is very sensible to the fitness function. Therefore, trying different and new energy functions is likely to enhance its accuracy. Finally, for improving the algorithm's performance, trying double core GPGPUs and different platforms such as OpenCL is suggested.

References

[1] A. Arranz, M. Alvar, J. Boal, A. S. Miralles, and A. de la Escalera. Genetic algorithm for stereo correspondence with a novel fitness function and occlusion handling. In *VISAPP (2)'13*, pages 294–299, 2013

[2] Y. Boykov, O. Veksler, and R. Zabih. Fast approximate energy minimization via graph cuts. *IEEE Transactions On Pattern Analysis And Machine Intelligence*, 23(11):1222–1239, 2001.

[3] D. Gallup, J. M. Frahm, P. Mordohai, Q. X. Yang, and M. Pollefeys. Real-time plane-sweeping stereo with multiple sweeping directions. *2007 IEEE Conference on Computer Vision and Pattern Recognition, Vols 1-8*, pages 2110–2117, 2007.

[4] S. Izadi, R. A. Newcombe, D. Kim, O. Hilliges, D. Molyneaux, S. Hodges, P. Kohli, J. Shotton, A. J. Davison, and A. Fitzgibbon. KinectFusion: real-time dynamic 3D surface reconstruction and interac- tion. In *ACM SIGGRAPH 2011 Talks*, SIGGRAPH '11, page 23:123:1, New York, NY, USA, 2011. ACM.

[5] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *Int.J.Comput.Vision*, 60(2):91–110, 2004.

[6] X. Mei, X. Sun, M. Zhou, S. Jiao, H. Wang, and X. Zhang. On building an accurate stereo matching system on graphics hardware. In *Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Confer- ence on*, pages 467 –474, nov. 2011.

[7] R. A. Newcombe, A. J. Davison, S. Izadi, P. Kohli, O. Hilliges, J. Shotton, D. Molyneaux, S. Hodges, D. Kim, and A. Fitzgibbon. KinectFusion: real-time dense surface mapping and tracking. In *Mixed and Augmented Reality (ISMAR), 2011 10th IEEE International Symposium on*, pages 127–136, 2011.

[8] D.-H. Nie, K.-P. Han, and H.-S. Lee. Stereo matching algorithm using population-based incremental learning on gpu. In *Intelligent Systems and Applications, 2009. ISA 2009. International Workshop on*, pages 1–4, 2009.

[9] S. Sinha, J.-M. Frahm, M. Pollefeys, and Y. Genc. Feature tracking and matching in video using programmable graphics hardware. *Machine Vision and Applications*, 2007.

[10] V. Vineet and P. J. Narayanan. Cuda cuts: Fast graph cuts on the gpu. In *Computer Vision and Pattern Recognition Workshops, 2008. CVPRW '08. IEEE Computer Society Conference on*, pages 1–8, 2008.

[11] R. Yang and M. Pollefeys. Multi-resolution real-time stereo on commodity graphics hardware, 2003.

[12] K. J. Yoon and I. S. Kweon. Adaptive support-weight approach for correspondence search. *Ieee Transactions On Pattern Analysis And Machine Intelligence*, 28(4):650–656, Apr 2006.

[13] Y. Zhao and G. Taubin. Real-time stereo on gpgpu using progressive multi-resolution adaptive windows. pages 420–432, 2011.

[14] Seoane, P., M. Gestal, and J. Dorado, "Approach for solving multimodal problems using Genetic Algorithms with Grouped into Species optimized with Predator-Prey", International Journal of Interactive Multimedia and Artificial Intelligence, vol. 1, no. 5, pp. 6-13, 06/2012

[15] Mey Rodríguez, M., and E. P. Gayoso, "EvoWild: a demosimulator about wild life", International Journal of Interactive Multimedia and Artificial Intelligence, vol. 1, issue Experimental Simulations, no. 1, pp. 25-30, 12/2008

Alvaro Arranz was born in Madrid in 1984. He obtained his Industrial Engineering degree, majoring in electronic engineering, from Universidad Pontificia Comillas-ICAI in 2007. Currently he is finishing his PhD on Computer Vision and working as a researcher and developer at Bitmonlab in Zed Worldwide. His research interests include Computer Vision, Robotics, Intelligent Systems and Human-Computer interaction.

Manuel Alvar was born in Málaga (Spain) in 1984. He obtained his Industrial Engineering degree, from Universidad Pontificia Comillas-ICAI; and the Engineering degree from the École Centrale de Paris (France) in 2007. His areas of interest are: Smartgrids, Mobile Autonomous Robots, Unmanned Helicopters, Computer Vision, Automatic Monitoring, Web Tools Development.

www.elasticbox.com
info@elasticbox.com
blog.elasticbox.com
Twitter: elasticbox

elasticbox

Deploy applications, not servers.

Challenges In Cloud Computing

Automation

- How long to deploy an application?
- What version do I use?
- How do I upgrade applications?

Portability

- How do I change providers?
- What is being used?
- How much does it cost?

Auto-Scaling

- Can my application auto-scale?
- How do I configure auto-scaling?

Disaster Recovery Planning

- Can my application tolerare faults?
- How do I recover my system?

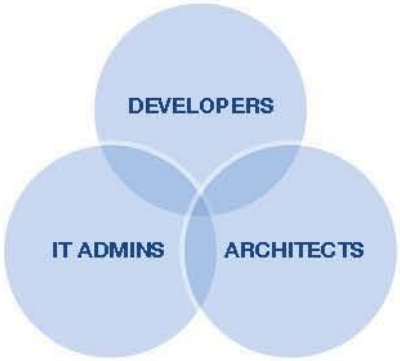
ElasticBox Solution

Automation

- ✓ Automatic Deployments
- ✓ Automatic Configuration
- ✓ Disaster Recovery

Runtime Environment

- ✓ Application Scaling
- ✓ Fault Tolerance
- ✓ Resource Clean-up
- ✓ Replication



Framework Design

- ✓ Architecture Policies
- ✓ Versioning
- ✓ Platform Management

Infrastructure Control

- ✓ Cost Analysis
- ✓ Policy Management
- ✓ Traceability

amazon web services™ · Windows Azure · hp · rackspace · openstack · EUCALYPTUS